OPT2024: 16th Annual Workshop on Optimization for Machine Learning

Tight Lower Bounds and Improved Convergence in Performative Prediction

Pedram Khorsandi[1,2] PEDRAM.KHORSANDI@MILA.QUEBEC
Rushil Gupta[1,2] RUSHIL.GUPTA@MILA.QUEBEC
Mehrnaz Mofakhami[1,2] MEHRNAZ.MOFAKHAMI@MILA.QUEBEC
Simon Lacoste-Julien*[1,2,3] SLACOSTE@UMONTREAL.CA
Gauthier Gidel*[1,2,3] GIDELGAU@MILA.QUEBEC
[1]*Mila - Quebec AI Institute*
[2] *Université de Montréal*
[3]*Canada CIFAR AI Chair*

Abstract

Performative prediction is a framework accounting for the shift in the data distribution induced by the prediction of a model deployed in the real world. Ensuring rapid convergence to a stable solution where the data distribution remains the same after the model deployment is crucial, especially in evolving environments. This paper extends the Repeated Risk Minimization (RRM) framework by utilizing historical datasets from previous retraining snapshots, yielding a class of algorithms that we call Affine Risk Minimizers and enabling convergence to a performatively stable point for a broader class of problems. We introduce a new upper bound for methods that use only the final iteration of the dataset and prove for the first time the tightness of both this new bound and the previous existing bounds within the same regime. We also prove that utilizing historical datasets can surpass the lower bound for last iterate RRM, and empirically observe faster convergence to the stable point on various performative prediction benchmarks. We offer at the same time the first lower bound analysis for RRM within the class of Affine Risk Minimizers, quantifying the potential improvements in convergence speed that could be achieved with other variants in our framework.

1. INTRODUCTION

Decision-making systems are increasingly integral to critical judgments in sectors such as public policy [5], healthcare [1], and education [19]. However, as these systems become more reliant on quantitative indicators, they become vulnerable to the effects described by Goodhart's Law: "When a measure becomes a target, it ceases to be a good measure" [7]. This principle is particularly relevant when predictive models not only forecast outcomes but also influence the behavior of individuals and organizations, leading to performative effects that can subvert the original goals of these systems.

Given these challenges, it is essential to develop predictive models that are not only accurate but also robust against the performative shifts they may provoke. The work by [21] addresses this challenge within the framework of *Repeated Risk Minimization (RRM)*, where they explore the dynamics of model retraining in the presence of performative feedback loops. In their approach, the authors propose an iterative method that adjusts the predictive model based on the distributional shifts caused by prior model deployments, aiming to stabilize the model performance despite the continuous evolution of the underlying data distribution. By characterizing the convergence

properties of their method, they provide a theoretical guarantee for the stability of the model at a performative equilibrium.

Our work extends this framework by leveraging the datasets collected at each snapshot during the retraining process, introducing a new class of algorithms called *Affine Risk Minimizers*. By utilizing historical data from previous updates, we show that it is possible to converge to a stable point for a broader class of problems that were previously unsolvable, extending beyond the bounds established in prior analyses [17]. We derive a new upper bound under less restrictive assumptions than [17] and provide the first tightness analysis for the framework in [21] as well as for our newly established rate. Our method, which incorporates historical datasets, demonstrates superior convergence properties both theoretically and experimentally.

Faster convergence is of particular importance in scenarios where the data distribution is subject to continuous change. By achieving more rapid convergence, our framework ensures that the model stabilizes more quickly, minimizing the period during which predictions may be unreliable.

Contributions. ① We establish a new upper bound, enhancing the convergence rate of RRM under less restrictive conditions; ② We establish the tightness of the analysis in both our framework and the framework proposed by [21]; ③ We introduce a new class of algorithms, named Affine Risk Minimizers, that provides convergence for a wider class of problems by utilizing linear combinations of datasets from earlier training snapshots; ④ We provide both theoretical and experimental enhancements, showcasing scenarios where this framework improves convergence; ⑤ Finally, we introduce the first technique for establishing theoretical lower bounds across each framework, detailing the maximum potential improvement in convergence rates achievable through the use of past datasets.

Code. `github.com/pedramkho/bounds_in_performative_prediction`

2. REPEATED RISK MINIMIZATION (RRM)

RRM iteratively retrains the model on the distribution it induces until it converges to a performatively stable classifier. Formally, consider a model with parameters $\theta \in \Theta$, and a distribution $D(\theta)$ that depends on these parameters. The performative risk is defined as:

$$\mathrm{PR}(\theta) = \mathbb{E}_{z\sim D(f_\theta)}\left[\ell(f_\theta(x), y)\right] \tag{1}$$

where $\ell(f_\theta(x), y)$ is the loss function for a data point $z = (x,\ y)$. A classifier is performatively stable if it minimizes the performative risk on the distribution it induces:

$$\theta_{\mathrm{PS}} = \arg\min_{\theta\in\Theta} \mathbb{E}_{z\sim D(f_{\theta_{\mathrm{PS}}})}\left[\ell(f_\theta(x), y)\right] \tag{2}$$

The RRM framework updates the model parameters by solving:

$$\theta^{t+1} = \arg\min_{\theta\in\Theta} \mathbb{E}_{z\sim D(f_{\theta^t})}\left[\ell(f_\theta(x), y))\right] \tag{3}$$

until convergence, i.e., $\theta^{t+1} \approx \theta^t$.

3. IMPROVED RATES AND OPTIMALITY OF ANALYSIS

Both Perdomo et al. [21] and Mofakhami et al. [17] derive convergence rates for RRM under distinct assumptions. The assumptions made in these studies reflect the sensitivity of the distribution map $\mathcal{D}(.)$ to changes in the model and the structural properties of the loss function. Specifically, Perdomo et al. [21] focuses on Wasserstein-based sensitivity and convexity with respect to the model parameters, while Mofakhami et al. [17] introduces a framework with Pearson χ^2-based sensitivity and strong convexity with respect to the predictions. Building on these foundations, and motivated by Mofakhami et al. [17] we now outline the assumptions for our framework:

Assumption 1 *ϵ-sensitivity w.r.t. Pearson χ^2 divergence: The distribution map $\mathcal{D}(f_\theta)$, with pdf p_{f_θ}, maintains ϵ-sensitivity with respect to Pearson χ^2 divergence. Formally, for any $f_\theta, f_{\theta'} \in \mathcal{F}$:*

$$\chi^2(\mathcal{D}(f_{\theta'}), \mathcal{D}(f_\theta)) \leq \epsilon \|f_\theta - f_{\theta'}\|_{f_\theta}^2, \tag{4}$$

where

$$\|f_\theta - f_{\theta'}\|_{f_{\theta*}}^2 := \int \|f_\theta(x) - f_{\theta'}(x)\|^2 p_{f_{\theta*}}(x) dx, \tag{5}$$

and

$$\chi^2(\mathcal{D}(f_{\theta'}), \mathcal{D}(f_\theta)) := \int \frac{\left(p_{f_{\theta'}}(z) - p_{f_\theta}(z)\right)^2}{p_{f_\theta}(z)} dz\,. \tag{6}$$

This assumption, inspired by prior work on Lipschitz continuity on $D(.)$, implies that if two models with similar prediction functions are deployed, the distributions they induce should also be similar.

Assumption 2 *Norm equivalency: The distribution map $\mathcal{D}(f_\theta)$ satisfies a bounded norm ratio with parameters $C \geq 1$ and $c \leq C$. For all $f_\theta, f_{\theta'} \in \mathcal{F}$:*

$$c\|f_\theta - f_{\theta'}\|_{f_{\theta^*}}^2 \leq \|f_\theta - f_{\theta'}\|^2 \leq C\|f_\theta - f_{\theta'}\|_{f_{\theta^*}}^2, \tag{7}$$

where

$$\|f_\theta - f_{\theta'}\|^2 = \int \|f_\theta(x) - f_{\theta'}(x)\|^2 p(x)\, dx, \tag{8}$$

and $p(x)$ is the initial distribution, referred to as the base distribution.

Assumption 3 *Strong convexity w.r.t. predictions: The loss function $\hat{y} \mapsto \ell(\hat{y}, y)$ is γ-strongly convex. Specifically, for all $y, \hat{y}_1, \hat{y}_2 \in \mathcal{Y}$:*

$$\ell(\hat{y}_1, y) \geq \ell(\hat{y}_2, y) + (\hat{y}_1 - \hat{y}_2)^\top \nabla_{\hat{y}} \ell(\hat{y}_2, y) + \frac{\gamma}{2}\|\hat{y}_1 - \hat{y}_2\|^2.$$

Assumption 4 *Bounded gradient norm: The loss function $\ell(f_\theta(x), y)$ has a bounded gradient norm, with an upper bound $M = \sup_{x,y,\theta} \|\nabla_{\hat{y}} \ell(f_\theta(x), y)\|$.*

Building upon Mofakhami et al. [17], we introduce a new theorem that demonstrates faster linear convergence for RRM, showing that stability can be achieved under less restrictive conditions.

Theorem 1 *Suppose the loss $\ell(f_\theta(x), y)$ is γ-strongly convex with respect to $f_\theta(x)$ (A3) and that the gradient norm with respect to $f_\theta(x)$ is bounded by $M = \sup_{x,y,\theta} \|\nabla_{\hat{y}} \ell(f_\theta(x), y)\|$ (A4). Let the distribution map $\mathcal{D}(\cdot)$ be ϵ-sensitive with respect to the Pearson χ^2 divergence (A1), satisfy a bounded norm ratio with parameters $C \geq 1$ and $c \leq C$ (A2), and the function space $\mathcal{F}$ be convex and compact under the norm $\|\cdot\|$.*

Then, for $G(\theta_t) = \arg\min_{\theta \in \Theta} \mathbb{E}_{z \sim \mathcal{D}(f_{\theta_t})} \ell(f_\theta(x), y)$, with $z = (x, y)$, we have[1]*:*

$$\|f_{G(\theta)} - f_{G(\theta')}\|_{f_\theta} \leq \frac{\sqrt{\epsilon} M}{\gamma} \|f_\theta - f_{\theta'}\|_{f_\theta}.$$

By the Schauder fixed-point theorem, a stable classifier $f_{\theta_{PS}}$ exists, and if $\frac{\sqrt{\epsilon} M}{\gamma} < 1$, RRM converges to a unique stable point $f_{\theta_{PS}}$ at a linear rate:

$$\|f_{\theta_t} - f_{\theta_{PS}}\|_{f_{\theta_{PS}}} \leq \left(\frac{\sqrt{\epsilon} M}{\gamma}\right)^t \|f_{\theta_0} - f_{\theta_{PS}}\|_{f_{\theta_{PS}}}.$$

This shows that RRM achieves linear convergence to a stable classifier, provided that $\frac{\sqrt{\epsilon} M}{\gamma} < 1$, ensuring that the mapping is contractive and guarantees convergence. The proof of this theorem is provided in Appendix D.

This result improves upon Mofakhami et al. [17] by eliminating the constant C from the rate, as defined in Assumption 2. Additionally, this approach can achieve improved rates of convergence, as discussed in Theorem 5, where we show how relaxing this assumption along with the new definition of ϵ-sensitivity leads to faster convergence.

Despite these improvements, the following theorem establishes a lower bound under the given assumptions, indicating that the convergence rate cannot be further improved without additional conditions:

Theorem 2 *Suppose that Assumptions 1-4 hold, with parameters ϵ, M, and γ such that $\frac{\sqrt{\epsilon} M}{\gamma} \leq 1$. Under these conditions, there exists a problem instance such that, utilizing RRM, the following holds:*

$$\|f_{\theta_t} - f_{\theta_{PS}}\|_{f_{\theta_{PS}}} = \Omega\left(\left(\frac{\sqrt{\epsilon} M}{\gamma}\right)^t\right). \tag{9}$$

If instead $\frac{\sqrt{\epsilon} M}{\gamma} > 1$, the bound is $\Omega(1)$, indicating non-convergence.

The full proof of this theorem can be found in Appendix F.

This result establishes the tightness of the convergence rate under the specific assumptions outlined earlier, demonstrating that the bound cannot be improved without imposing more restrictive assumptions. A similar tightness analysis for the framework proposed by Perdomo et al. [21] is provided in Appendix C.

Our theorems show that given the assumptions in either framework, the convergence rate for RRM reaches a fundamental lower bound. This implies that further improvements in convergence speed would require either more restrictive assumptions or a novel optimization framework.

In the next section, we present a new approach that surpasses this lower bound by leveraging data from previous training snapshots, allowing for faster convergence beyond the limits established by existing methods.

1. Throughout this work, whenever we refer to $f_{G(\theta)}$, it denotes $f_{\hat{\theta}}$, where $\hat{\theta} \in G(\theta)$.

4. USAGE OF OLD SNAPSHOTS

Our method introduces an alternative approach to improve convergence. Instead of relying solely on the current data distribution induced by $D(f_{\theta^t})$, we leverage datasets from previous training snapshots $\{D(f_{\theta^i})\}_{i=0}^{t-1}$. The updated framework optimizes model parameters over an aggregated distribution:

$$\theta^{t+1} = \arg\min_{\theta\in\Theta} \mathbb{E}_{(x,\,y)\sim D_t} \left[\ell(f_\theta(x),\, y)\right] \tag{10}$$

where D_t is an affine combination of previous distributions, formulated as:

$$D_t = \sum_{i=0}^{t-1} \alpha_i^{(t)} D(f_{\theta^i}), \quad \text{s.t.} \quad \sum_{i=0}^{t-1} \alpha_i^{(t)} = 1 \tag{11}$$

We refer to this class of algorithms as *Affine Risk Minimizers*. As demonstrated in Appendix **??** (Lemma 8), the set of stable points for this class of algorithms coincides with those obtained through standard RRM.

The following theorem provides theoretical evidence of improved convergence, which will be further supported by experiments in Section A.

Although this improvement is based on a more restrictive set of assumptions, these assumptions still hold for the lower bound described in Theorem 2.

Lemma 1 *Consider the class of problems for which Assumptions 1-4 are satisfied, and let the distribution map $\mathcal{D}(.)$ be $\frac{\epsilon}{C}$-sensitive with respect to the base distribution within the convex function space $\mathcal{F}$. Formally, for any $f_\theta, f_{\theta'} \in \mathcal{F}$,*

$$\chi^2(\mathcal{D}(f_{\theta'}), \mathcal{D}(f_\theta)) \leq \frac{\epsilon}{C}\|f_\theta - f_{\theta'}\|^2,$$

where $\|f_\theta - f_{\theta'}\|^2 := \int \|f_\theta(x) - f_{\theta'}(x)\|^2 p(x)dx$. The distribution at iteration t is given by

$$D_t = \frac{1}{2}D(f_{\theta^t}) + \frac{1}{2}D(f_{\theta^{t-1}}).$$

Under these conditions, the following convergence property holds for the iterative sequence generated by Equation 10:

$$\|f_{\theta^{t+1}} - f_{\theta^t}\| = \left(\sqrt{\tfrac{\sqrt{3}+2}{4}}\frac{\sqrt{\epsilon}M}{\gamma}\right) m_t, \tag{12}$$

where $m_t = \max\{\|f_{\theta^t} - f_{\theta^{t-1}}\|,\, \|f_{\theta^{t-1}} - f_{\theta^{t-2}}\|\}$.

The problem class defined here aligns with that in Theorem 2 for the case $C \approx 1$.

Theorem 3 *If $\left(\sqrt{\frac{\sqrt{3}+2}{4}}\frac{\sqrt{\epsilon}M}{\gamma}\right) < 1$, the sequence described in Lemma 1 forms a Cauchy sequence, converging to a stable point.*

This theorem shows that in the regime where

$$\frac{\sqrt{\epsilon}M}{\gamma} < \frac{1}{\sqrt{\frac{\sqrt{3}+2}{4}}} \approx 1.035$$

and $C \approx 1$, it is possible to converge to a stable point despite the lower bound in Theorem 2, while operating under the same conditions as the lower bound. A detailed proof of this theorem, along with Lemma 1 is provide'd in Appendix G, where we also establish that the algorithm generates a Cauchy sequence, thereby guaranteeing convergence to the stable point.

In the next section, we explore the lower bound for the convergence rates achievable using any affine combination of previous snapshots.

5. LOWER BOUNDS FOR AFFINE RISK MINIMIZERS

In the previous section, we established the potential for convergence across a wider class of problems using Affine Risk Minimizers. This prompts the question of how much the convergence speed can be improved, which we address in this section.

We propose the first distinct lower bounds for the framework described in Section 3 and that of Perdomo et al. [21] for the class of Affine Risk Minimizers. The lower bound for our framework is presented in the following section, while the corresponding result for Perdomo et al. [21] is detailed in Section C.2.

Theorem 4 *Suppose that Assumptions 1-4 hold. Then, there exists a problem instance in this regime, and for any algorithm in the Affine Risk Minimizers class, such that:*

$$\|f_{\theta^t} - f_{\theta_{PS}}\|_{f_{\theta_{PS}}} = \Omega\left(\left(\frac{1}{\frac{1}{e}+2}\frac{\sqrt{\epsilon}M}{\gamma}\right)^t\right). \tag{13}$$

This demonstrates that the convergence rate for the class of problems satisfying Assumptions 1-4 cannot exceed the given lower bound.

CONCLUSION

This work introduces Affine Risk Minimizers, a novel class of algorithms that utilize historical datasets from retraining snapshots to address convergence challenges in performative prediction. By deriving a new upper bound and demonstrating its tightness, along with the first lower bound analysis for this class of methods, we establish theoretical and empirical evidence for improved convergence under less restrictive assumptions. These results highlight the advantages of aggregating past data, enabling the resolution of a broader range of performative prediction problems while improving model stability in evolving environments.

ACKNOWLEDGMENTS

This research was supported by the Canada CIFAR AI Chair Program and the NSERC Discovery Grant RGPIN-2023-04373. We gratefully acknowledge the computational resources provided by Calcul Quebec (calculquebec.ca) and the Digital Research Alliance of Canada (alliancecan.

ca), which enabled part of the experiments. Simon Lacoste-Julien is a CIFAR Associate Fellow in the Learning in Machines & Brains program. We also thank António Góis and Alan Milligan for their valuable feedback, which notably contributed to this work.

References

[1] Gwyn Bevan and Christopher Hood. What's measured is what matters: Targets and gaming in the english public health care system. *Public Administration*, 2006. doi: 10.1111/j.1467-9299.2006.00600.x. https://doi.org/10.1111/j.1467-9299.2006.00600.x.

[2] Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *AISTATS*, 2022. https://proceedings.mlr.press/v151/brown22a.html.

[3] Roy Dong, Heling Zhang, and Lillian Ratliff. Approximate regions of attraction in learning with decision-dependent distributions. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *AISTATS*, 2023. https://proceedings.mlr.press/v206/dong23b.html.

[4] D.C. Dowson and B.V. Landau. The fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982. https://www.sciencedirect.com/science/article/pii/0047259X8290077X.

[5] Michael Fire and Carlos Guestrin. Over-optimization of academic publishing metrics: observing Goodhart's Law in action. *GigaScience*, 8(6):giz053, 2019. https://doi.org/10.1093/gigascience/giz053.

[6] Ziv Goldfeld, Rami Pellumbi, and Kia Khezeli. Lecture 6: f-divergences. In *ECE 5630: Information Theory for Data Transmission, Security and Machine Learning*, 2020. https://people.ece.cornell.edu/zivg/ECE_5630_Lectures6.pdf.

[7] C. A. E. Goodhart. *Problems of Monetary Management: The UK Experience*, pages 91–121. Macmillan Education UK, 1984. https://doi.org/10.1007/978-1-349-17295-5_4.

[8] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *ITCS*. Association for Computing Machinery, 2016. https://doi.org/10.1145/2840728.2840730.

[9] Zachary Izzo, Lexing Ying, and James Zou. How to learn when data reacts to your model: Performative gradient descent. In Marina Meila and Tong Zhang, editors, *ICML*, 2021. https://proceedings.mlr.press/v139/izzo21a.html.

[10] Meena Jagadeesan, Celestine Mendler-Dünner, and Moritz Hardt. Alternative microfoundations for strategic classification. In Marina Meila and Tong Zhang, editors, *ICML*, 2021. https://proceedings.mlr.press/v139/jagadeesan21a.html.

[11] Meena Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner. Regret minimization with performative feedback. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *ICML*, 2022. https://proceedings.mlr.press/v162/jagadeesan22a.html.

[12] Qiang Li and Hoi-To Wai. State dependent performative prediction with stochastic approximation. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *AISTATS*, 2022. https://proceedings.mlr.press/v151/li22c.html.

[13] Licong Lin and Tijana Zrnic. Plug-in performative optimization. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *ICML*, 2024. https://proceedings.mlr.press/v235/lin24ab.html.

[14] Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *NeurIPS*, 2020. https://proceedings.neurips.cc/paper_files/paper/2020/file/33e75ff09dd601bbe69f351039152189-Paper.pdf.

[15] Celestine Mendler-Dünner, Frances Ding, and Yixin Wang. Anticipating performativity by predicting from predictions. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *NeurIPS*. Curran Associates, Inc., 2022. https://arxiv.org/abs/2208.07331.

[16] John P Miller, Juan C Perdomo, and Tijana Zrnic. Outside the echo chamber: Optimizing the performative risk. In Marina Meila and Tong Zhang, editors, *ICML*, 2021. https://proceedings.mlr.press/v139/miller21a.html.

[17] Mehrnaz Mofakhami, Ioannis Mitliagkas, and Gauthier Gidel. Performative prediction with neural networks. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *AISTATS*, volume 206, 2023. https://proceedings.mlr.press/v206/mofakhami23a.html.

[18] Adhyyan Narang, Evan Faulkner, Dmitriy Drusvyatskiy, Maryam Fazel, and Lillian J. Ratliff. Multiplayer performative prediction: learning in decision-dependent games. *JMLR*, 2024. https://www.jmlr.org/papers/volume24/narang23a/narang23a.pdf.

[19] Sharon L. Nichols and David C. Berliner. *Collateral damage: How high-stakes testing corrupts America's schools*. Harvard Education Press, 2007. https://psycnet.apa.org/record/2007-03254-000.

[20] Frank Nielsen and Kazuki Okamura. On the f-divergences between densities of a multivariate location or scale family. *Statistics and Computing*, 2024. ISSN 1573-1375. doi: 10.1007/s11222-023-10373-6. https://doi.org/10.1007/s11222-023-10373-6.

[21] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In Hal Daumé III and Aarti Singh, editors, *ICML*, 2020. https://proceedings.mlr.press/v119/perdomo20a.html.

[22] Ben Rank, Stelios Triantafyllou, Debmalya Mandal, and Goran Radanovic. Performative reinforcement learning in gradually shifting environments, 2024. https://arxiv.org/abs/2402.09838.

Appendix A. Experiments

We conduct experiments in two semi-synthetic environments to evaluate whether aggregating past snapshots improves convergence to the performatively stable point. We present an empirical comparison of different aggregation windows for prior snapshots. At each time step t, we form D_t by aggregating the datasets from the training snapshots as

$$D_t = \frac{1}{\tau} \sum_{i=t-\tau+1}^{t} D(f_{\theta^i}),$$

where we compare methods using various values of τ, including $\tau = 1, 2, 4, \frac{t}{2}$, and 'all' (which includes all snapshots up to time t).

We first begin with a discussion on our evaluation metric, followed by detailed case-studies on both the credit scoring environment [17] and the rideshare markets [18] in subsequent subsections (A.1, A.2).

Evaluation Metric. Throughout our experiments, we focus on changes in loss as an effect of performativity. We define $\Delta\mathcal{R}_t$, i.e. the loss shift due to performativity at time t, as the absolute difference in loss observed by a model before and after the data distribution has changed due to performative effects while keeping the model's state constant.

$$\Delta\mathcal{R}_t = \left|\mathbb{E}_{z\sim\mathcal{D}(\theta^t)}[\ell(z,\, \theta^t)] - \mathbb{E}_{z\sim\mathcal{D}(\theta^{t-1})}[\ell(z,\, \theta^t)]\right| \tag{14}$$

This metric allows for clearer comparisons between methods by minimizing overlap in the plots, unlike the performative risk (see Eq.1).

A.1. Credit Scoring

Setup. Inspired by [17], we use the *Resample-if-Rejected (RIR)* procedure to model distribution shifts in a controlled experimental setting. This methodology involves users strategically altering their data to influence the classification outcome.

Let us consider a base distribution with probability density function p and a function $g : f_\theta(x) \mapsto g(f_\theta(x))$ indicating the probability of rejection based on the prediction $f_\theta(x) \in \mathbb{R}$. The modified distribution p_{f_θ}, under the *RIR* mechanism, evolves as follows:

- Sample x from p.
- With probability $1 - g(f_\theta(x))$, accept and output x. Otherwise, resample from p.

Our data comes from Kaggle's *Give Me Some Credit* dataset[2], which includes features $x \in \mathbb{R}^{11}$ and labels $y \in \{0, 1\}$, where $y = 1$ indicates a defaulting applicant. We partition the features into two sets: strategic and non-strategic. We assume independence between strategic and non-strategic features. While non-strategic features remain fixed, the strategic features are resampled using the *RIR* procedure with a rejection probability $g(f_\theta(x)) = f_\theta(x) + \delta$. We use a scaled sigmoid function after the second layer. This scales $f_\theta(x)$ to the interval $[0, 1 - \delta]$, ensuring that $g(f_\theta(x)) \in [\delta, 1]$ remains a valid probability. Further implementation details are available in Appendix K.

2. Give me Some Credit Dataset, 2011: https://www.kaggle.com/c/GiveMeSomeCredit

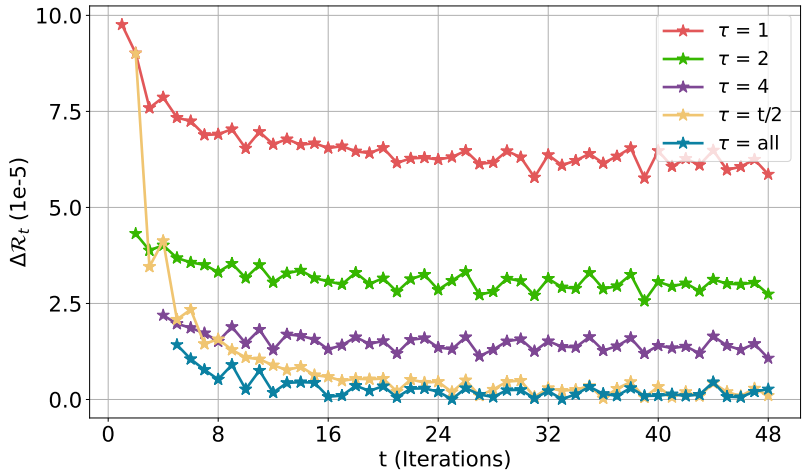

Figure 1: Loss shift due to performativity for the credit scoring environment. To accurately measure Performative Risk we average over 500 runs per method. Increasing the size of aggregation window τ from $1 \to 2 \to 4 \to t/2 \to all$ reduces the loss shifts and hence, reaches the stable point faster.

Theorem 5 *Let $f_\theta(x) \in [0, 1-\delta]$ for all $\theta \in \Theta$, where $0 < \delta < 1$ is fixed. Then, for $g(f_\theta(x)) = f_\theta(x) + \delta$, RIR is ϵ-sensitive as defined in Assumption 1 with $\epsilon = \mathcal{O}(\delta^{-\frac{3}{2}})$.*

This result provides an example where our rate surpasses the rate previously derived in [17] ($\mathcal{O}(\delta^{-2})$ within the same framework). Furthermore, for any value of M and γ, our rate can guarantee convergence for a wider class of problems. The proof of this theorem, along with justifications for the improved rate, is presented in Appendix J.

Results. The outcomes of this case study are shown in Figure 1. For larger window sizes (τ), we omit the initial iterations in the figure because they follow the same update rule as smaller τ methods, leading to identical values.
Figure 1 demonstrates the advantage of using older snapshots in the optimization process. As the window size increases from 1 to 2, we observe a near-half reduction in the loss shift, particularly in the early iterations, with the improvement persisting even after 50 iterations. While larger windows continue to reduce the loss shift, the marginal gains decrease as window size increases. This is evident from the similarity between the curves for window sizes $t/2$, and 'all'.
The decreasing marginal gains elicit a trade-off against the time, memory, and resource consumption. As the window size increases, both time per iteration and the memory consumption increase linearly. Thus, the user has to pick the right aggregation window τ based on the application and the resources available to achieve the desired convergence speed while respecting the logistical constraints. The corresponding performative risk plot can also be found in Appendix K.

A.2. Revenue Maximization in Ride Share Market

Setup. This is a two-player semi-synthetic game between two ride-share providers, Uber and Lyft, both trying to maximize their respective revenues. Each player takes an action in this game by setting their price for the riders across 11 different locations in the same city of Boston, MA. The price set by one firm directly influences the demand observed by both firms. The demand constitutes

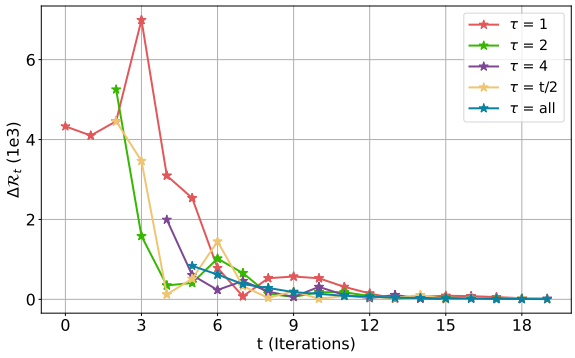

Figure 2: Loss shift due to performativity

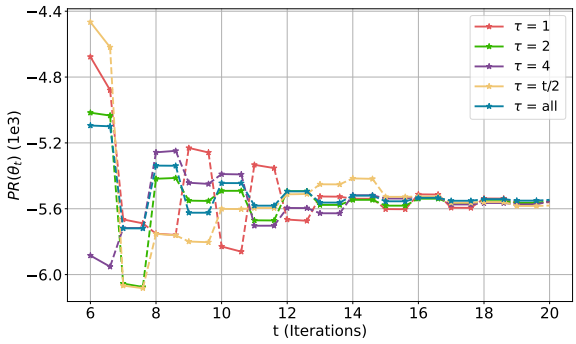

Figure 3: Performative Risk

Figure 4: The plot shows loss shift due to performativity and performative risk across the iterations for player 1 in the game between two firms. The values in the plot are means over 200 runs. Increasing the aggregation window size τ leads to lower loss shifts even in this simple game and hence, faster convergence than just relying on the dataset from the current timestamp.

the data distribution and at each time step, a total of 25 demand samples are sampled for a firm i and the optimal response is found by minimizing Equation 16 for a maximum of 40 re-training steps. The simulations use the publicly available *Uber and Lyft dataset from Boston, MA* on Kaggle[3].

Notations and Equations. Let $i = 1, 2$ denote the two firms in the game. Inspired by [18], each firm i observes a demand z_i that depends linearly on the firm's price $\mathbf{x_i}$ and its opponent's price $\mathbf{x_{-i}}$ as follows:

$$\mathbf{z}_i = \mathbf{A}_i \mathbf{x}_i + \mathbf{A}_{-i} \mathbf{x}_{-i} + \xi, \quad \xi \sim \mathcal{N}(\mathbf{z}_{\text{base}}, 1) \tag{15}$$

where $\mathbf{z_{base}}$ is the mean demand observed at each of the 11 locations, as measured in the kaggle dataset. Each demand sample is a vector of dimension 11.
$\mathbf{A_i}$ and $\mathbf{A_{-i}}$ are fixed matrices representing the price elasticity of demand, i.e. the change in demand due to a unit change in price for the player i and $-i$ (opponent) respectively. We introduce interactions between the ride prices in a location and the demand in a different location within the same city by making $\mathbf{A}$ matrices non-diagonal. Additionally, note that the price elasticities $\mathbf{A_i}$ will always be negative as the firm will experience less demand if it increases its price. Similarly, the price elasticities $\mathbf{A_{-i}}$ will be positive.
Each player observes a revenue of $\mathbf{z_i^T x_i}$. Thus, the loss function that each player i seeks to minimize in the RRM framework can be described as:

$$\mathbf{x}_i^{t+1} = \arg\min_{\mathbf{x}_i} \mathbb{E}_{z_i \sim \mathcal{D}_t} \left[-\mathbf{z}_i^T \mathbf{x}_i + \frac{\lambda}{2} \|\mathbf{x}_i\|^2 \right] \tag{16}$$

where λ is a hyperparameter for the regularization term (= 70 for our experiments). For any player i, each element of $\mathbf{x_i}$ is clipped to be between the range of $[-30, 30]$ and the initial price $\mathbf{x_i^0}$ is sampled randomly from a uniform distribution on $[0, 1]$.

3. Uber and Lyft dataset from Boston, MA, 2019: https://www.kaggle.com/datasets/brllrb/uber-and-lyft-dataset-boston-ma

Results. Figure 4 shows the plot for loss shift due to performativity and the performative risk versus the iterations averaged over 200 runs. For this plot, we assume player 1 is the player who makes the predictions and adjusts to the performative effects introduced due to the actions of player 2. It can be clearly observed that as we increase the aggregation window from $1 \rightarrow 2 \rightarrow 4 \rightarrow t/2 \rightarrow all$, we get mostly lower loss shifts and hence, an improvement in the convergence rate. Since we start at random price value, taking the past into account in the beginning makes the algorithm worse but the effect is neutralized as the data from more time steps is observed. Given the simple linear nature of the problem, this is a significant improvement and provides evidence for our claims about using the data from the previous snapshots.
Secondly, performative risk plot in figure 4 also highlights that all methods converge to points having very close values of performative risk, with the methods having larger τ showing oscillations with smaller amplitude.

Appendix B. Auxiliary Lemmas and Technical Results

Lemma 2 *(Expectation of a Gaussian-Weighted Exponential Function) Let* $\mathbf{x} \sim N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$*. Then the expected value of* $\mathbf{x} \exp\left(-\frac{1}{2e}\|\mathbf{x}\|^2\right)$ *is given by:*

$$\mathbb{E}\left[\mathbf{x} \exp\left(-\frac{1}{2e}\|\mathbf{x}\|^2\right)\right] = \exp\left(-\frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2}\left(1 - \frac{1}{\sigma^2\left(\frac{1}{e} + \frac{1}{\sigma^2}\right)}\right)\right) \cdot \frac{\boldsymbol{\mu}}{\sigma^2\left(\frac{1}{e} + \frac{1}{\sigma^2}\right)}.$$

Proof: The expected value is expressed as:

$$\mathbb{E}\left[\mathbf{x} \exp\left(-\frac{1}{2e}\|\mathbf{x}\|^2\right)\right] = \int_{\mathbb{R}^n} \mathbf{x} \exp\left(-\frac{1}{2e}\|\mathbf{x}\|^2\right) \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x} - \boldsymbol{\mu}\|^2\right) d\mathbf{x}.$$

Merging the exponentials:

$$\exp\left(-\frac{1}{2}\left(\frac{1}{e} + \frac{1}{\sigma^2}\right)\|\mathbf{x}\|^2 + \frac{1}{\sigma^2}\mathbf{x}^T\boldsymbol{\mu}\right) \cdot \exp\left(-\frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2}\right).$$

Completing the square yields:

$$\exp\left(-\frac{1}{2}\left(\frac{1}{e} + \frac{1}{\sigma^2}\right)\left\|\mathbf{x} - \frac{\boldsymbol{\mu}/\sigma^2}{\frac{1}{e} + \frac{1}{\sigma^2}}\right\|^2\right) \cdot \exp\left(-\frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2}\left(1 - \frac{1}{\sigma^2\left(\frac{1}{e} + \frac{1}{\sigma^2}\right)}\right)\right)$$

Since the integral is over a Gaussian distribution with mean $\frac{\boldsymbol{\mu}/\sigma^2}{\frac{1}{e}+\frac{1}{\sigma^2}}$, after multiplying by the constant term, we obtain:

$$\mathbb{E}[\mathbf{x} \exp\left(-\frac{1}{2e}\|\mathbf{x}\|^2\right)] = \exp\left(-\frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2}\left(1 - \frac{1}{\sigma^2\left(\frac{1}{e} + \frac{1}{\sigma^2}\right)}\right)\right) \cdot \frac{\boldsymbol{\mu}}{\sigma^2\left(\frac{1}{e} + \frac{1}{\sigma^2}\right)}.$$

Lemma 3 *(Bound on Chi-Square Divergence for Convex Combinations) Let* P_1 *and* P_2 *be probability distributions on* $\mathbb{R}$*, and let* Q *be a reference distribution. For any* $\alpha \in [0, 1]$ *and any* $a > 0$*, the following inequality holds:*

$$\chi^2(\alpha P_1 + (1-\alpha)P_2, Q) \leq (1+a)\alpha^2\chi^2(P_1, Q) + \left(1 + \frac{1}{a}\right)(1-\alpha)^2\chi^2(P_2, Q).$$

Proof: We begin by expanding the chi-square divergence using its definition, followed by applying Young's inequality.

$$\begin{aligned}
\chi^2(\alpha P_1 + (1-\alpha)P_2, Q) &= \int_{-\infty}^{\infty} \frac{(\alpha p_1(x) + (1-\alpha)p_2(x) - q(x))^2}{q(x)}\, dx \\
&= \int_{-\infty}^{\infty} \frac{(\alpha(p_1(x) - q(x)) + (1-\alpha)(p_2(x) - q(x)))^2}{q(x)}\, dx \\
&\leq \int_{-\infty}^{\infty} \left[(1+a)\alpha^2 \frac{(p_1(x) - q(x))^2}{q(x)} + \left(1 + \frac{1}{a}\right)(1-\alpha)^2 \frac{(p_2(x) - q(x))^2}{q(x)}\right] dx \\
&\quad \text{(by Young's inequality)} \\
&= (1+a)\alpha^2\chi^2(P_1, Q) + \left(1 + \frac{1}{a}\right)(1-\alpha)^2\chi^2(P_2, Q).
\end{aligned} \tag{17}$$

Lemma 4 *(Inverse of an antisymmetric of a Jordan Normal Form Matrix) Let $A \in \mathbb{R}^{d\times d}$ be defined as:*

$$A = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 0 & 1 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 1 \end{bmatrix},$$

and let $bI - cA$ be an invertible matrix where $\frac{c}{b} \leq \frac{1}{2}$ and A is as defined above. Then the inverse of $(bI - cA)$ applied to e_1, the first standard basis vector, has the following form for large d:

$$\mathbf{v} = (bI - cA)^{-1}\frac{e_1}{L} = \frac{1}{cL}\begin{bmatrix} (\frac{b}{c} - 1)^{-1} \\ (\frac{b}{c} - 1)^{-2} \\ (\frac{b}{c} - 1)^{-3} \\ \vdots \\ (\frac{b}{c} - 1)^{-d} \end{bmatrix}.$$

Moreover, the sum below is:

$$\sum_{i=t}^{d} \mathbf{v}_i = \Omega\left(\left(\frac{c}{b}\right)^t\right),$$

for $d \geq 2T$ when T is large, and $t \leq T$.

Proof: The matrix A has the following form:

$$A = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 0 & 1 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 1 \end{bmatrix}.$$

Thus, $(bI - cA)$ takes the form:

$$bI - cA = c\begin{bmatrix} \frac{b}{c} - 1 & 0 & 0 & \dots & 0 \\ -1 & \frac{b}{c} - 1 & 0 & \dots & 0 \\ 0 & -1 & \frac{b}{c} - 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & -1 & \frac{b}{c} - 1 \end{bmatrix}.$$

We continue by computing the inverse of the lower triangular matrix with diagonal entries $\lambda_1, \lambda_2, \dots, \lambda_d$ and subdiagonal entries of -1 as shown below:

$$\begin{bmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ -1 & \lambda_2 & 0 & \dots & 0 \\ 0 & -1 & \lambda_3 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & -1 & \lambda_d \end{bmatrix}^{-1} = \begin{bmatrix} \lambda_1^{-1} & 0 & 0 & \dots & 0 \\ \lambda_1^{-1}\lambda_2^{-1} & \lambda_2^{-1} & 0 & \dots & 0 \\ \lambda_1^{-1}\lambda_2^{-1}\lambda_3^{-1} & \lambda_2^{-1}\lambda_3^{-1} & \lambda_3^{-1} & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \lambda_1^{-1}\lambda_2^{-1}\dots\lambda_d^{-1} & & \dots & \lambda_{d-1}^{-1}\lambda_d^{-1} & \lambda_d^{-1} \end{bmatrix}.$$

Using the formula above (diagonal entries $\lambda_1 = \frac{b}{c} - 1, \lambda_2 = \frac{b}{c} - 1, \ldots, \lambda_d = \frac{b}{c} - 1$ and subdiagonal entries of -1) the inverse of $bI - cA$ will have the form:

$$\frac{1}{c}\begin{bmatrix} (\frac{b}{c}-1)^{-1} & 0 & 0 & \dots & 0 \\ (\frac{b}{c}-1)^{-2} & (\frac{b}{c}-1)^{-1} & 0 & \dots & 0 \\ (\frac{b}{c}-1)^{-3} & (\frac{b}{c}-1)^{-2} & (\frac{b}{c}-1)^{-1} & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ (\frac{b}{c}-1)^{-d} & & \dots & (\frac{b}{c}-1)^{-2} & (\frac{b}{c}-1)^{-1} \end{bmatrix}.$$

Now, applying this inverse to the vector $\frac{e_1}{L}$, where $e_1 = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}^T$, we get the following:

$$\mathbf{v} = (bI - cA)^{-1}\frac{e_1}{L} = \frac{1}{cL}\begin{bmatrix} (\frac{b}{c}-1)^{-1} \\ (\frac{b}{c}-1)^{-2} \\ (\frac{b}{c}-1)^{-3} \\ \vdots \\ (\frac{b}{c}-1)^{-d} \end{bmatrix}.$$

Sum of the entries from index t to d is:

$$\sum_{i=t}^{d} \mathbf{v}_i = \frac{1}{cL}\left((\frac{b}{c}-1)^{-t} + (\frac{b}{c}-1)^{-t-1} + \cdots + (\frac{b}{c}-1)^{-d}\right).$$

This is a geometric series. The closed form of the sum is:

$$\sum_{i=t}^{d} \mathbf{v}_i = \frac{1}{cL} \cdot (\frac{b}{c}-1)^{-t} \cdot \frac{1 - (\frac{b}{c}-1)^{-(d-t+1)}}{1 - (\frac{b}{c}-1)^{-1}}.$$

For large $d \geq 2t$ and $\frac{b}{c} - 1 \geq 1$, this sum can be approximated by the leading term:

$$\sum_{i=t}^{d} \mathbf{v}_i \approx \frac{1}{cL} \cdot (\frac{b}{c}-1)^{-t} \cdot \frac{1}{1 - (\frac{b}{c}-1)^{-1}}.$$

Thus, applying the inequality $\frac{1}{\frac{1}{x}-1} \leq x$ for all $x < 1$, we obtain the following lower bound for the sum:

$$\sum_{i=t}^{d} \mathbf{v}_i = \Omega\left(\frac{1}{cL} \cdot \left(\frac{c}{b}\right)^t\right).$$

Lemma 5 *Let $\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$ be two multivariate normal distributions with means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^d$ and covariance matrices $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d \times d}$. The squared 1-Wasserstein distance between these distributions is bounded by:*

$$W_1^2(\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1), \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)) \leq \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 + tr\left(\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1\Sigma_2\right)^{1/2}\right).$$

This expression bounds the Wasserstein distance between two multivariate normal distributions, as shown in Dowson and Landau [4].

Lemma 6 *Let $\mathcal{N}(\boldsymbol{\mu}_1, \Sigma)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \Sigma)$ be two multivariate normal distributions with means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^d$ and a shared covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. The χ^2-divergence between these distributions is bounded by:*

$$\chi^2(\mathcal{N}(\boldsymbol{\mu}_1, \Sigma), \mathcal{N}(\boldsymbol{\mu}_2, \Sigma)) = 1 - e^{-\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)} \leq \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

This provides the χ^2-divergence between two multivariate normal distributions, as shown in Nielsen and Okamura [20].

Lemma 7 *Any projection proj$(.)$ from $\mathbb{R}^d$ into any convex set $\mathcal{C} \in \mathbb{R}^d$ is a continuous function.*

Proof: To prove that the projection is continuous, we need to show that if $x_n \to x$ in $\mathbb{R}^d$, then $\text{proj}_{\mathcal{C}}(x_n) \to \text{proj}_{\mathcal{C}}(x)$.

Let $y_n = \text{proj}_{\mathcal{C}}(x_n)$ and $y = \text{proj}_{\mathcal{C}}(x)$. Since $y_n \in \mathcal{C}$ and y_n minimizes the distance to x_n, we have:

$$\|x_n - y_n\| \leq \|x_n - y\| \quad \text{for all } n.$$

As $x_n \to x$, the right-hand side $\|x_n - y\| \to \|x - y\|$, and thus $\|x_n - y_n\|$ is bounded. Since the sequence $\{y_n\}$ is bounded and lies in the compact set $\mathcal{C}$, it has a convergent subsequence $y_{n_k} \to \bar{y} \in \mathcal{C}$. By the continuity of the distance function, we have:

$$\|x - \bar{y}\| = \lim_{k\to\infty} \|x_{n_k} - y_{n_k}\|.$$

As $y = \text{proj}_{\mathcal{C}}(x)$ minimizes the distance from x to $\mathcal{C}$, it follows that $\bar{y} = y$, and thus $y_n \to y$. Therefore, $\text{proj}_{\mathcal{C}}(x_n) \to \text{proj}_{\mathcal{C}}(x)$, proving continuity.

Lemma 8 *The set of stable points for any method in the class of Affine Risk Minimizers is equivalent to the set of stable points for standard RRM.*

Proof: Consider the mapping for an affine risk minimizer using the last τ iterates, defined as:

$$G_\tau(\theta^{t-1}, \theta^{t-2}, \ldots, \theta^{t-\tau}) = (\theta^t, \theta^{t-1}, \ldots, \theta^{t-\tau+1}),$$

where

$$\theta^t = \arg\min_{\theta\in\Theta} \mathbb{E}_{z\sim D_t}[\ell(f_\theta(x), y)].$$

At a stable point, the mapping satisfies:

$$(\theta^t, \theta^{t-1}, \ldots, \theta^{t-\tau}) = (\theta^t, \theta^{t-1}, \ldots, \theta^{t-\tau+1}),$$

which implies that:

$$\theta^t = \theta^{t-1} = \cdots = \theta^{t-\tau}.$$

We now show that every stable point for this mapping is also a stable point for the standard RRM mapping, defined as:

$$G(\theta^{t-1}) = \theta^t.$$

From the definition of D_t, we have:

$$D_t = \sum_{i=t-\tau}^{t-1} \alpha_i^{(t)} D(\theta^i) = D(\theta^{t-1}),$$

since $\sum_{i=t-\tau}^{t-1} \alpha_i^{(t)} = 1$. Therefore:

$$\theta^t = \arg\min_{\theta\in\Theta} \mathbb{E}_{z\sim D_t}[\ell(f_\theta(x), y)] = \arg\min_{\theta\in\Theta} \mathbb{E}_{z\sim D(\theta^{t-1})}[\ell(f_\theta(x), y)] = G(\theta^{t-1}),$$

implying that any stable point for G_τ is also a stable point for G.

Conversely, if $\theta^t = \theta^{t-1}$ at a stable point of G, then iterating the mapping G_τ τ times yields the sequence:

$$\theta^t = \theta^{t+1} = \cdots = \theta^{t+\tau},$$

A similar argument shows that this stable point satisfies:

$$\theta^{t+\tau} = \arg\min_{\theta\in\Theta} \mathbb{E}_{z\sim D(\theta^{t+\tau-1})}[\ell(f_\theta(x), y)] = \arg\min_{\theta\in\Theta} \mathbb{E}_{z\sim D_{t+\tau}}[\ell(f_\theta(x), y)],$$

because:

$$D_{t+\tau} = \sum_{i=t}^{t+\tau-1} \alpha_i^{(t)} D(\theta^i) = D(\theta^{t-1}).$$

Which leads to,

$$G_\tau(\theta^t, \theta^{t+1}, \ldots, \theta^{t+\tau}) = (\theta^{t-1}, \theta^t, \ldots, \theta^{t+\tau-1})$$

showing that this stable point is also stable for G_τ.

Thus, the set of stable points is equivalent for both mappings.

Lemma 9 *Let $a, b > 0$ with $b \leq 4a$. For any integer $t \geq 0$, the mixed power term $b^{\lceil t/2\rceil} a^{\lfloor t/2\rfloor}$ is upper-bounded as follows:*

$$b^{\lceil t/2\rceil} a^{\lfloor t/2\rfloor} \leq 2(ab)^{t/2}.$$

Proof: We consider two cases based on the parity of t:

1. **Case t even:** If t is even, then $\lceil t/2\rceil = \lfloor t/2\rfloor = t/2$. Thus,

$$b^{\lceil t/2\rceil} a^{\lfloor t/2\rfloor} = b^{t/2} a^{t/2} = (ab)^{t/2} < 2(ab)^{t/2}.$$

2. **Case t odd:** If t is odd, then $\lceil t/2\rceil = \frac{t+1}{2}$ and $\lfloor t/2\rfloor = \frac{t-1}{2}$. Therefore,

$$b^{\lceil t/2\rceil} a^{\lfloor t/2\rfloor} = b^{(t+1)/2} a^{(t-1)/2} = (ab)^{t/2} \cdot \left(\frac{b}{a}\right)^{1/2}.$$

Since $b \leq 4a$, it follows that $\left(\frac{b}{a}\right)^{1/2} \leq 2$, so

$$b^{\lceil t/2\rceil} a^{\lfloor t/2\rfloor} \leq 2(ab)^{t/2}.$$

Combining both cases, we conclude that:

$$b^{\lceil t/2\rceil} a^{\lfloor t/2\rfloor} \leq 2(ab)^{t/2}.$$

Lemma 10 *Let A_1 and A_2 be two probability distributions and let B_1 and B_2 be another two probability distributions. Then, we have*

$$\chi^2\left(\frac{A_1+A_2}{2}, \frac{B_1+B_2}{2}\right) \leq \chi^2(A_1, B_1) + \chi^2(A_2, B_2).$$

Proof:

To compute the χ^2 divergence between the averages $\frac{A_1+A_2}{2}$ and $\frac{B_1+B_2}{2}$, we start with the definition:

$$\chi^2\left(\frac{A_1+A_2}{2}, \frac{B_1+B_2}{2}\right) = \int_{-\infty}^{\infty} \frac{\left(\frac{p_1^A(x)+p_2^A(x)}{2} - \frac{p_1^B(x)+p_2^B(x)}{2}\right)^2}{\frac{p_1^B(x)+p_2^B(x)}{2}}\, dx.$$

Simplifying the numerator, we get:

$$= \frac{1}{2}\int_{-\infty}^{\infty} \frac{\left(p_1^A(x) + p_2^A(x) - p_1^B(x) - p_2^B(x)\right)^2}{p_1^B(x) + p_2^B(x)}\, dx.$$

Applying the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we can further bound this as follows:

$$\leq \frac{1}{2}\int_{-\infty}^{\infty} \frac{2(p_1^A(x) - p_1^B(x))^2 + 2(p_2^A(x) - p_2^B(x))^2}{p_1^B(x) + p_2^B(x)}\, dx.$$

By distributing the terms, this becomes:

$$= \int_{-\infty}^{\infty} \frac{(p_1^A(x) - p_1^B(x))^2}{p_1^B(x) + p_2^B(x)} + \frac{(p_2^A(x) - p_2^B(x))^2}{p_1^B(x) + p_2^B(x)}\, dx.$$

Now, since $\frac{1}{p_1^B(x)+p_2^B(x)} \leq \frac{1}{p_1^B(x)}$ and $\frac{1}{p_1^B(x)+p_2^B(x)} \leq \frac{1}{p_2^B(x)}$, we can split the integral as follows:

$$\leq \int_{-\infty}^{\infty} \frac{(p_1^A(x) - p_1^B(x))^2}{p_1^B(x)} + \frac{(p_2^A(x) - p_2^B(x))^2}{p_2^B(x)}\, dx.$$

By definition of the χ^2 divergence, this final expression is equivalent to:

$$= \chi^2(A_1, B_1) + \chi^2(A_2, B_2).$$

Thus, we have shown that

$$\chi^2\left(\frac{A_1+A_2}{2}, \frac{B_1+B_2}{2}\right) \leq \chi^2(A_1, B_1) + \chi^2(A_2, B_2),$$

which completes the proof.

Lemma 11 *Let $\eta = f_{G(\theta')} - f_{G(\theta)}$. Suppose the function space $\mathcal{F}$ is convex, and*

$$G(\theta) = \arg\min_{\theta' \in \Theta} \mathbb{E}_z \left[\ell(f_{\theta'}(x), y)\right],$$

where $z = (x, y) \sim p_{f_\theta}$ represents the distribution induced by the model f_θ, and ℓ is a differentiable loss function. Then the following inequality holds:

$$\int \eta(x)^\top \nabla_y \ell(f_{G(\theta)}(x), y)\, p_{f_\theta}(z)\, dz \geq 0.$$

Refer to Mofakhami et al. [17] for the proof.

Appendix C. Lower Bounds for Perdomo et al. [21]

C.1. Tightness Analysis in Perdomo et al. [21]'s Framework

In their work, Perdomo et al. [21] make a set of assumptions that differs from Assumption 1-4. Their ϵ-sensitivity assumption is with respect to the Wasserstein distance, and their strong convexity assumption is with respect to the parameters. Formally they make the following set of assumptions to show the convergence of RRM

Assumption 5 *The distribution map $\theta \mapsto D(\theta)$ is ϵ-sensitive w.r.t $\mathcal{W}_1$:*

$$\mathcal{W}_1(\mathcal{D}(\theta), \mathcal{D}(\theta')) \leq \epsilon \|\theta - \theta'\|_2,$$

the loss function $\theta \mapsto \ell(z : \theta)$ of the performative risk (1) is γ-strongly convex for any $z \in \mathcal{Z}$ and $z \mapsto \nabla_z \ell(z : \theta)$ is β-Lipschitz for any $\theta \in \Theta$.

Under these assumptions and for $\frac{\beta\epsilon}{\gamma} < 1$, Perdomo et al. [21] showed that RRM does converge to a performatively stable point at a rate:[4]

$$\|\theta_t - \theta_{PS}\| \leq \left(\frac{\beta\epsilon}{\gamma}\right)^t \|\theta_0 - \theta_{PS}\|. \tag{18}$$

Theorem 6 *There exists a problem instance and an initialization θ_0 following assumptions 5 such that employing RRM, we have:*

$$\|\theta^t - \theta_{PS}\| = \Omega\left(\left(\frac{\epsilon\beta}{\gamma}\right)^t \|\theta_0 - \theta_{PS}\|\right). \tag{19}$$

The proof of this result is provided in Appendix E.

C.2. Lower Bound with Perdomo et al. [21]'s Assumption

We show that the convergence rate for RRM provided in Equation 18 is optimal among the class of *Affine Risk Minimizers* up to a factor 2.

Theorem 7 *There exists a problem instance and an initialization θ_0 following Assumption 5 such that for any algorithm in the Affine Risk Minimizers class, we have:*

$$\|\theta^t - \theta_{PS}\| = \Omega\left(\left(\frac{\epsilon\beta}{2\gamma}\right)^t \|\theta_0 - \theta_{PS}\|\right). \tag{20}$$

The proof of this result can be found in Appendix H.

To further illustrate this, Figure 5 provides empirical evidence supporting the theoretical lower bound derived for Perdomo et al. [21]. The figure shows the convergence of $\|\theta - \theta_{\text{PS}}\|$ over multiple iterations for various combinations of previous snapshots. As indicated by the dotted line, the lower bound is never violated, demonstrating that the theoretical result holds in practice. The experimental setup for these results is also detailed in Appendix H.

4. Note that if $\frac{\beta\epsilon}{\gamma} \geq 1$ the convergence rate is vacuous. In that case, a performatively stable point may not even exist.

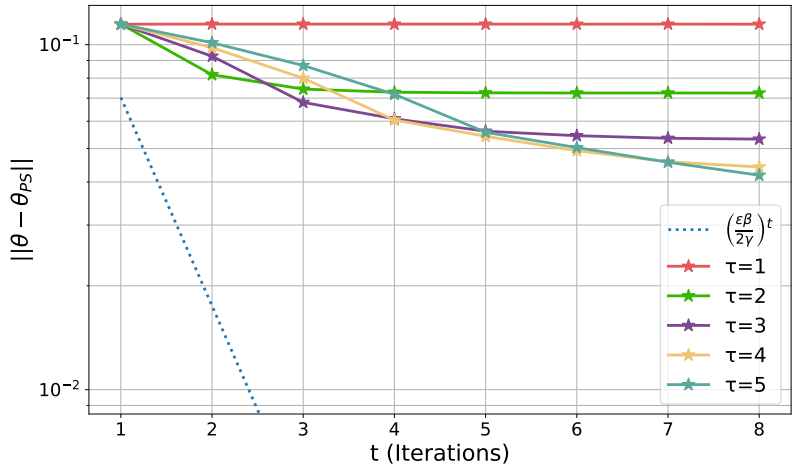

Figure 5: Convergence of $\|\theta^t - \theta_{\text{PS}}\|$ over iterations t, for different values of τ, where τ defines the aggregation of datasets from training snapshots, i.e. $D_t = \sum_{i=t-\tau+1}^{t} \frac{1}{\tau} D(\theta_i)$. The dotted line represents our lower bound derived for Perdomo et al. [21], with $\epsilon = 2.49$, $\beta = 1$, and $\gamma = 5.0$. The experiment follows the setup across all methods and demonstrates the validity of this lower bound by showing that $\|\theta^t - \theta_{\text{PS}}\|$ does not decay below the lower bound, providing experimental evidence for our theoretical result.

Appendix D. Proof of Theorem 1

The proof of Theorem 1 largely follows the approach in Mofakhami et al. [17], with some modifications to remove the need for the bounded norm ratio assumption. To facilitate readability, we have restated the common parts from the proof in Mofakhami et al. [17].

Fix θ and θ' in Θ. Let $h : \mathcal{F} \mapsto \mathbb{R}$ and $h' : \mathcal{F} \mapsto \mathbb{R}$ be two functionals defined as follows:

$$h(f_{\hat{\theta}}) = E_{z\sim\mathcal{D}(f_\theta)}[\ell(f_{\hat{\theta}}(x), y)] = \int \ell(f_{\hat{\theta}}(x), y) p_{f_\theta}(z) dz \tag{21}$$

$$h'(f_{\hat{\theta}}) = E_{z\sim\mathcal{D}(f_{\theta'})}[\ell(f_{\hat{\theta}}(x), y)] = \int \ell(f_{\hat{\theta}}(x), y) p_{f_{\theta'}}(z) dz \tag{22}$$

where each data point z is a pair of features x and label y.

For a fixed $z = (x, y)$, due to strong convexity of $\ell(f_\theta(x), y)$ in $f_\theta(x)$ we have:

$$\begin{aligned}\ell(f_{G(\theta)}(x), y) - \ell(f_{G(\theta')}(x), y) \geq \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right)^\top \nabla_{\hat{y}} \ell(f_{G(\theta')}(x), y) \\ + \frac{\gamma}{2} \|f_{G(\theta)}(x) - f_{G(\theta')}(x)\|^2.\end{aligned} \tag{23}$$

Now take integral over z, and define $\|f_{G(\theta)} - f_{G(\theta')}\|_{f_\theta}^2 = \int \|f_{G(\theta)}(x) - f_{G(\theta')}(x)\|^2 p_{f_\theta}(z) dz$:

$$\begin{aligned}h(f_{G(\theta)}) - h(f_{G(\theta')}) \geq \left(\int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right)^\top \nabla_{\hat{y}} \ell(f_{G(\theta')}(x), y) p_{f_\theta}(z) dz \right) \\ + \frac{\gamma}{2} \|f_{G(\theta)} - f_{G(\theta')}\|_{f_\theta}^2.\end{aligned} \tag{24}$$

Similarly:

$$h(f_{G(\theta')}) - h(f_{G(\theta)}) \geq \left(\int \left(f_{G(\theta')}(x) - f_{G(\theta)}(x)\right)^\top \nabla_{\hat{y}} \ell(f_{G(\theta)}(x), y) p_{f_\theta}(z) dz \right) + \frac{\gamma}{2} \|f_{G(\theta)} - f_{G(\theta')}\|_{f_\theta}^2. \tag{25}$$

Since $f_{G(\theta)}$ minimizes h, the following result can be achieved through the convexity of the function space,(Lemma 11):

$$\int \left(f_{G(\theta')}(x) - f_{G(\theta)}(x)\right)^\top \nabla_{\hat{y}} \ell(f_{G(\theta)}(x), y) p_{f_\theta}(z) dz \geq 0. \tag{26}$$

Adding (24) and (25) and using the above inequality, we conclude:

$$-\gamma \|f_{G(\theta)} - f_{G(\theta')}\|_{f_\theta}^2 \geq \int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right)^\top \nabla_{\hat{y}} \ell(f_{G(\theta')}(x), y) p_{f_\theta}(z) dz. \tag{27}$$

This is a key inequality that will be used later in the proof.

Now recall that there exists M such that $M = \sup_{x,y,\theta} \|\nabla_{\hat{y}} \ell(f_\theta(x), y)\|$ and the distribution map over data is ϵ-sensitive w.r.t Pearson χ^2 divergence, i.e.

$$\chi^2(\mathcal{D}(f_{\theta'}), \mathcal{D}(f_\theta)) \leq \epsilon \|f_\theta - f_{\theta'}\|_{f_\theta}^2. \tag{28}$$

With this in mind, we do the following calculations:

$$\begin{aligned}
&\left| \int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right)^\top \nabla_{\hat{y}} \ell(f_{G(\theta')}(x), y) p_{f_\theta}(z) dz - \int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right)^\top \nabla_{\hat{y}} \ell(f_{G(\theta')}(x), y) p_{f_{\theta'}}(z) dz \right| \\
&= \left| \int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right)^\top \nabla_{\hat{y}} \ell(f_{G(\theta')}(x), y) \left(p_{f_\theta}(z) - p_{f_{\theta'}}(z)\right) dz \right| \\
&\overset{(*)}{\leq} \int \left| \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right)^\top \nabla_{\hat{y}} \ell(f_{G(\theta')}(x), y) \left(p_{f_\theta}(z) - p_{f_{\theta'}}(z)\right) \right| dz \\
&\leq M \int \left| \|f_{G(\theta)}(x) - f_{G(\theta')}(x)\| \left(p_{f_\theta}(z) - p_{f_{\theta'}}(z)\right) \right| dz \\
&= M \int \left| \|f_{G(\theta)}(x) - f_{G(\theta')}(x)\| \frac{p_{f_\theta}(z) - p_{f_{\theta'}}(z)}{p_{f_\theta}(z)} p_{f_\theta}(z) \right| dz \\
&= M \left| \int \left| \|f_{G(\theta)}(x) - f_{G(\theta')}(x)\| \frac{p_{f_\theta}(z) - p_{f_{\theta'}}(z)}{p_{f_\theta}(z)} \right| p_{f_\theta}(z) dz \right| \\
&\overset{\text{Cauchy-Schwarz Ineq.}}{\leq} M \left(\int \|f_{G(\theta)}(x) - f_{G(\theta')}(x)\|^2 p_{f_\theta}(z) dz \right)^{\frac{1}{2}} \left(\int \left(\frac{p_{f_\theta}(z) - p_{f_{\theta'}}(z)}{p_{f_\theta}(z)} \right)^2 p_{f_\theta}(z) dz \right)^{\frac{1}{2}} \\
&= M \|f_{G(\theta)} - f_{G(\theta')}\|_{f_\theta} \sqrt{\chi^2(\mathcal{D}(f_{\theta'}), \mathcal{D}(f_\theta))}
\end{aligned}$$

$(*)$ comes from the fact that $\left|\int f(x)dx\right| \leq \int |f(x)|dx$, and the Cauchy-Schwarz inequality states that $|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$.

We conclude from the above derivations that:

$$\begin{aligned}
&\left| \int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right)^\top \nabla_{\hat{y}} \ell(f_{G(\theta')}(x), y) p_{f_\theta}(z) dz - \int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right)^\top \nabla_{\hat{y}} \ell(f_{G(\theta')}(x), y) p_{f_{\theta'}}(z) dz \right| \\
&\leq M \|f_{G(\theta)} - f_{G(\theta')}\|_{f_\theta} \sqrt{\chi^2(\mathcal{D}(f_{\theta'}), \mathcal{D}(f_\theta))}.
\end{aligned} \tag{29}$$

Similar to inequality (26), since $f_{G(\theta')}$ minimizes h', one can prove:

$$\int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right)^\top \nabla_{\hat{y}} \ell(f_{G(\theta')}(x), y) p_{f_{\theta'}}(z) dz \geq 0. \tag{30}$$

From (27) we know that $\int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right)^\top \nabla_{\hat{y}} \ell(f_{G(\theta')}(x), y) p_{f_\theta}(z) dz$ is negative, so with this fact alongside (29) and (30), we can write:

$$\int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right)^\top \nabla_{\hat{y}} \ell(f_{G(\theta')}(x), y) p_{f_\theta}(z) dz \geq -M \|f_{G(\theta)} - f_{G(\theta')}\|_{f_\theta} \sqrt{\chi^2(\mathcal{D}(f_{\theta'}), \mathcal{D}(f_\theta))}. \tag{31}$$

Combining (27) and (31), we obtain:

$$\begin{aligned} &\gamma \|f_{G(\theta)} - f_{G(\theta')}\|_{f_\theta}^2 \leq M \|f_{G(\theta)} - f_{G(\theta')}\|_{f_\theta} \sqrt{\chi^2(\mathcal{D}(f_{\theta'}), \mathcal{D}(f_\theta))} \\ &\Rightarrow \|f_{G(\theta)} - f_{G(\theta')}\|_{f_\theta} \leq \frac{M}{\gamma} \sqrt{\chi^2(\mathcal{D}(f_{\theta'}), \mathcal{D}(f_\theta))} \overset{(28)}{\leq} \frac{\sqrt{\epsilon} M}{\gamma} \|f_\theta - f_{\theta'}\|_{f_\theta} \end{aligned} \tag{32}$$

To prove the existence of a fixed point, we use the Schauder fixed point theorem. Define

$$\mathcal{U} : f \in \mathcal{F} \to \arg\min_{f' \in \mathcal{F}} \mathop{\mathbb{E}}_{z \sim \mathcal{D}(f)} \ell(f'(x), y).$$

For this function, $\mathcal{U}(f_\theta) = f_{G(\theta)}$. So instead of Equation 32, we can write:

$$\|\mathcal{U}(f_\theta) - \mathcal{U}(f_{\theta'})\|_{f_\theta} \leq \frac{\sqrt{\epsilon} M}{\gamma} \|f_\theta - f_{\theta'}\|_{f_\theta}. \tag{33}$$

Using Assumption 2, we derive the following bound,

$$\|\mathcal{U}(f_\theta) - \mathcal{U}(f_{\theta'})\| \leq \left(\sqrt{\frac{C}{c}}\right) \frac{\sqrt{\epsilon} M}{\gamma} \|f_\theta - f_{\theta'}\|. \tag{34}$$

This inequality shows that for any $f_{\theta_0} \in \mathcal{F}$, if $\lim_{n\to\infty} \|f_{\theta_n} - f_\theta\| = 0$, then $\lim_{n\to\infty} \|\mathcal{U}(f_{\theta_n}) - \mathcal{U}(f_\theta)\| = 0$, which proves the continuity of $\mathcal{U}$ with respect to the norm $\|.\|$. Thus, since $\mathcal{U}$ is a continuous function from the convex and compact set $\mathcal{F}$ to itself, the Schauder fixed point theorem ensures that $\mathcal{U}$ has a fixed point. Therefore, $f_{\theta_{\text{PS}}}$ exists such that $f_{G(\theta_{\text{PS}})} = f_{\theta_{\text{PS}}}$.

If we set $\theta = \theta_{\text{PS}}$ and $\theta' = \theta_{t-1}$ for θ_{PS} being any sample in the set of stable classifiers, we know that $G(\theta) = \theta_{\text{PS}}$ and $G(\theta') = \theta_t$. So we will have:

$$\|f_{\theta_t} - f_{\theta_{\text{PS}}}\|_{f_{\theta_{\text{PS}}}} \leq \frac{\sqrt{\epsilon} M}{\gamma} \|f_{\theta_{t-1}} - f_{\theta_{\text{PS}}}\|_{f_{\theta_{\text{PS}}}}. \tag{35}$$

Thus,

$$\|f_{\theta_t} - f_{\theta_{\text{PS}}}\|_{f_{\theta_{\text{PS}}}} \leq \frac{\sqrt{\epsilon} M}{\gamma} \|f_{\theta_{t-1}} - f_{\theta_{\text{PS}}}\|_{f_{\theta_{\text{PS}}}} \leq \left(\frac{\sqrt{\epsilon} M}{\gamma}\right)^t \|f_{\theta_0} - f_{\theta_{\text{PS}}}\|_{f_{\theta_{\text{PS}}}}. \tag{36}$$

Note that Equation 36 applies to any stable point. Suppose there are two distinct stable points, $f_{\theta^1_{PS}}$ and $f_{\theta^2_{PS}}$. By the definition of stable points and using Equation 33, we have:

$$\|\mathcal{U}(f_{\theta^1_{PS}}) - \mathcal{U}(f_{\theta^2_{PS}})\| = \|f_{\theta^1_{PS}} - f_{\theta^2_{PS}}\|_{f_{\theta^1_{PS}}} \leq \frac{\sqrt{\epsilon}M}{\gamma}\|f_{\theta^1_{PS}} - f_{\theta^2_{PS}}\|_{f_{\theta^1_{PS}}}.$$

Under the assumption that $\frac{\sqrt{\epsilon}M}{\gamma} < 1$, the inequality above ensures that $f_{\theta^1_{PS}} = f_{\theta^2_{PS}}$ [5] and the stable point must be unique. Thus, Equation 36 confirms that RRM converges to a unique stable classifier at a linear rate.

5. It is important to clarify that $f_{\theta^1_{PS}} = f_{\theta^2_{PS}}$ does not imply $\forall x\ f_{\theta^1_{PS}}(x) = f_{\theta^2_{PS}}(x)$. Instead, it indicates that $\|f_{\theta^1_{PS}} - f_{\theta^2_{PS}}\| = 0$.

Appendix E. Proof of Theorem 6

In this section, we examine the tightness of the analysis presented in Perdomo et al. [21] by considering a specific loss function and designing a particular performativity framework. We focus on the loss function $\ell(z, \theta) = \frac{\gamma}{2}\|\theta - \frac{\beta}{\gamma}z\|^2$, which is γ-strongly convex with respect to the parameter θ and its gradient w.r.t. θ is β-Lipschitz, aligning with the assumptions stipulated in Perdomo et al. [21].

We model performativity through the following distribution: $z \sim \mathcal{N}(\epsilon\theta, \sigma^2)$. According to Lemma 5 the 1-Wasserstein distance between two normal distributions is upper bounded by:

$$W_1(\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)) \leq \sqrt{(\mu_1 - \mu_2)^2} = \epsilon\|\theta_1 - \theta_2\|$$

it follows that the distribution mapping specified is ϵ-sensitive, as described in Perdomo et al. [21].

Under these conditions, the RRM process results in the following update mechanism:

$$\theta^{t+1} = \epsilon\frac{\beta}{\gamma}\theta^t = (\epsilon\frac{\beta}{\gamma})^t\theta^0$$

This arises because:

$$\begin{aligned}\theta^{t+1} &= \arg\min_\theta \mathbb{E}_{z\sim\mathcal{D}(\theta^t)}[\ell(z,\theta)] = \arg\min_\theta \mathbb{E}_{z\sim\mathcal{D}(\theta^t)}\left[\frac{\gamma}{2}\theta^2 - \beta\theta z + \frac{\beta^2}{2\gamma}z^2\right] \\ &= \arg\min_\theta \mathbb{E}_{z\sim\mathcal{D}(\theta^t)}\left[\frac{\gamma}{2}\theta^2 - \beta\theta z\right] = \arg\min_\theta \frac{\gamma}{2}\theta^2 - \beta\epsilon\theta\theta^t = \epsilon\frac{\beta}{\gamma}\theta^t\end{aligned}$$

This progression directly corresponds to the upper bound suggested by Perdomo et al. [21], confirming that the analysis is tight. No further refinement of the analytical model would mean a faster convergence rate for the given set of assumptions as detailed in Perdomo et al. [21].

Appendix F. Proof of Theorem 2

We define the model fitting function as $f_\theta(x) = \theta$, and the corresponding loss function is:

$$\ell(x, \theta) = \frac{1}{2\gamma} \left\| \gamma f_\theta(x) - M \operatorname{proj}_{\|.\|=0.95}(x) \right\|^2,$$

where $\operatorname{proj}_{\|.\|=0.95}$ denotes the projection onto the surface of a ball with radius 0.95. By setting $\theta \in \Theta = \{z \mid \|z\| \leq 0.05 \min\{\frac{M}{\gamma}, \frac{1}{\sqrt{\epsilon}}\}\}$, we ensure that the gradient norm remains smaller than M. Since the loss function is γ-strongly convex, it satisfies both Assumptions 4 and 3.

Throughout this proof $\|\theta_1 - \theta_2\| = \|f_{\theta_1} - f_{\theta_2}\|_{f_{\theta'}}$ for any choice of θ'.

We define the distribution mapping as follows:

$$D(\theta) = N\left(\sqrt{\epsilon}\theta, \frac{1}{2}\right),$$

The χ^2-divergence between two distributions $D(\theta_1) = N(\mu_1, \sigma)$ and $D(\theta_2) = N(\mu_2, \sigma)$, where $\mu_1 = \sqrt{\epsilon}\theta_1$ and $\mu_2 = \sqrt{\epsilon}\theta_2$, with $\sigma = \frac{1}{2}$, is given by (Lemma 6):

$$\chi^2(N(\mu_1, \Sigma), N(\mu_2, \Sigma)) \leq \frac{1}{2}(\mu_1 - \mu_2)^\top \Sigma^{-1}(\mu_1 - \mu_2) = \epsilon\|\theta_1 - \theta_2\|^2 = \epsilon\|f_{\theta_1} - f_{\theta_2}\|^2_{f_{\theta_1}}.$$

Thus, the χ^2-divergence between the distributions is bounded by $\epsilon\|\theta_1 - \theta_2\|^2$, making it ϵ-sensitive according to Assumption 1. Note that

With this set up one would derive the update rule:

$$\theta^{t+1} = \operatorname{proj}_\Theta\left(\frac{M}{\gamma}\mathbb{E}\left[\operatorname{proj}(x)\right]\right) = \operatorname{proj}_\Theta\left(\frac{M}{\gamma}\operatorname{erf}\left(\frac{2\mathbb{E}\left[x\right]}{\sqrt{2}}\right)\right)$$

Using,

$$\operatorname{erf}\left(\frac{2x}{\sqrt{2}}\right) \geq x \quad \forall x \leq 0.05,$$

and given that $\mathbb{E}[x] = \sqrt{\epsilon}\theta \leq 0.05 \min\left\{\frac{\sqrt{\epsilon}M}{\gamma}, 1\right\}$ by the definition of Θ, the condition holds.

$$\theta^{t+1} \geq \operatorname{proj}_\Theta\left(\frac{M}{\gamma}\mathbb{E}\left[x\right]\right) = \operatorname{proj}_\Theta\left(\frac{M\sqrt{\epsilon}}{\gamma}\theta^t\right)$$

Assuming we start with θ^0 in the feasible set and operate in the regime where $\frac{M\sqrt{\epsilon}}{\gamma} \leq 1$, the projection into the feasible set can be omitted. Therefore, we have:

$$\theta^t \geq \left(\frac{M\sqrt{\epsilon}}{\gamma}\right)^t \theta^0.$$

It is clear that $\theta = 0$ is the stable point in this setup, so:

$$\|\theta^t - \theta_{PS}\| = \Omega\left(\left(\frac{M\sqrt{\epsilon}}{\gamma}\right)^t\right).$$

In other words:

$$\|f_{\theta^t} - f_{\theta_{PS}}\|_{f_{\theta_{PS}}} = \Omega\left(\left(\frac{M\sqrt{\epsilon}}{\gamma}\right)^t\right).$$

For the case where $\frac{M\sqrt{\epsilon}}{\gamma} > 1$, the projection remains constrained to the surface of the ball Θ, preventing convergence to the stable point.

Appendix G. Proof of Lemma 1 and Theorem 3

This proof is heavily inspired by the proof of Theorem 1 in Appendix D. We start by prescribing stronger assumptions that imply this paper's set of assumptions.

Assumption 6 *ϵ-sensitivity with respect to Pearson χ^2 divergence (version 2): The distribution map $\mathcal{D}(f_\theta)$ maintains ϵ-sensitivity with respect to Pearson χ^2 divergence. For all $f_\theta, f_{\theta'} \in \mathcal{F}$:*

$$\chi^2(\mathcal{D}(f_{\theta'}), \mathcal{D}(f_\theta)) \leq \frac{\epsilon}{C}\|f_\theta - f_{\theta'}\|^2, \tag{37}$$

where $\|f_\theta - f_{\theta'}\|^2$ is defined in Equation 8.

Note that, combining Assumptions 2 and 6, we can infer Assumption 1:

$$\chi^2(\mathcal{D}(f_{\theta'}), \mathcal{D}(f_\theta)) \leq \frac{\epsilon}{C}\|f_\theta - f_{\theta'}\|^2 \leq \epsilon\|f_\theta - f_{\theta'}\|^2_{f_{\theta*}}.$$

Following the methodology described for Theorem 2 in Mofakhami et al. [17], we begin by defining the functional evaluations at consecutive time steps as follows:

$$h^t(f_{\hat{\theta}}) = \mathbb{E}_{z\sim\mathcal{D}_t}[\ell(f_{\hat{\theta}}, z)] = \int \ell(f_{\hat{\theta}}, z)p_t(z)\,dz,$$
$$h^{t-1}(f_{\hat{\theta}}) = \mathbb{E}_{z\sim\mathcal{D}_{t-1}}[\ell(f_{\hat{\theta}}, z)] = \int \ell(f_{\hat{\theta}}, z)p_{t-1}(z)\,dz,$$

where $p_t(z)$ denotes the probability density function of sample z from the distribution $\mathcal{D}_t$.

Utilizing the convexity of ℓ and Lemma 1 from Mofakhami et al. [17], following the line of argument in equation 17 of Mofakhami et al. [17], we establish the following inequality:

$$-\gamma\|f_{\theta^{t+1}} - f_{\theta^t}\|^2_{p_t} \geq \int \left(f_{\theta^{t+1}}(x) - f_{\theta^t}(x)\right)^\top \nabla_{\hat{y}}\ell(f_{\theta^t}(x), y)p_t(z)\,dz, \tag{38}$$

where $\|f_{\theta^{t+1}} - f_{\theta^t}\|^2_{p_t}$ represents the squared norm, calculated as:

$$\|f_{\theta^{t+1}} - f_{\theta^t}\|^2_{p_t} = \int \|f_{\theta^{t+1}}(x) - f_{\theta^t}(x)\|^2 p_t(z)\,dz.$$

and $p_t(x) = \frac{1}{2}p_{f_{\theta^t}}(x) + \frac{1}{2}p_{f_{\theta^{t-1}}}(x)$, Using the bounded gradient assumption, we deduce:

$$\int \left(f_{\theta^{t+1}}(x) - f_{\theta^t}(x)\right)^\top \nabla_{\hat{y}}\ell(f_{\theta^t}(x), y)p_t(z)\,dz \geq -M\|f_{\theta^{t+1}} - f_{\theta^t}\|_{p_t}\sqrt{\chi^2(\mathcal{D}_t, \mathcal{D}_{t-1})}. \tag{39}$$

Now combining equations 38 and 39 we get,

$$\gamma\|f_{\theta^{t+1}} - f_{\theta^t}\|_{p_t} \leq M\sqrt{\chi^2(\mathcal{D}_t, \mathcal{D}_{t-1})}. \tag{40}$$

Note that Equation 40, is a direct consequence of Assumptions 3-4, 6, and doesn't rely on definition of $\mathcal{D}_t$ (refer to Equation 32 for the proof). In other words, if we define our method as the mapping

$$\mathcal{U}(f_{\theta_1},\, f_{\theta_2}) = \arg\min_{f\in\mathcal{F}} \mathbb{E}_{(x,\,y)\sim D(f_{\theta_1},\, f_{\theta_2})}\left[\ell(f(x),\, y)\right],$$

where $D(f_{\theta_1},\, f_{\theta_2}) = \frac{D(f_{\theta_1})+D(f_{\theta_2})}{2}$, then,

$$\gamma\|\mathcal{U}(f_{\theta_1},\, f_{\theta_2}) - \mathcal{U}(f_{\theta_1'},\, f_{\theta_2'})\|_{p_d} \leq M\sqrt{\chi^2(D(f_{\theta_1},\, f_{\theta_2}),\, D(f_{\theta_1'},\, f_{\theta_2'}))}, \tag{41}$$

where, p_d is probability density function of distribution $D(f_{\theta_1},\, f_{\theta_2})$. We use this information further on in the proof.

The remaining task is to bound the χ^2 divergence as follows:

$$\begin{aligned}
\chi^2(\mathcal{D}_{t-1}, \mathcal{D}_t) &\leq (1+a)\alpha^2\chi^2(D(f_{\theta^{t-1}}), \alpha D(f_{\theta^t}) + (1-\alpha)D(f_{\theta^{t-1}})) \\
&\quad + \left(1+\frac{1}{a}\right)(1-\alpha)^2\chi^2(D(f_{\theta^{t-2}}), \alpha D(f_{\theta^t}) + (1-\alpha)D(f_{\theta^{t-1}})) \\
&\quad \text{(by Lemma 3)} \\
&\leq (1+a)\alpha^3\chi^2(D(f_{\theta^{t-1}}), D(f_{\theta^t})) \\
&\quad + \left(1+\frac{1}{a}\right)(1-\alpha)^2\alpha\chi^2(D(f_{\theta^{t-2}}), D(f_{\theta^t})) \\
&\quad + \left(1+\frac{1}{a}\right)(1-\alpha)^3\chi^2(D(f_{\theta^{t-2}}), D(f_{\theta^{t-1}})) \\
&\quad \text{(by Proposition 6.1 of Goldfeld et al. [6], convexity of } f\text{-divergence with respect to its arguments)} \\
&\leq \frac{\epsilon}{C}(1+a)\alpha^3\|f_{\theta^{t-1}} - f_{\theta^t}\|^2 \\
&\quad + \frac{\epsilon}{C}\left(1+\frac{1}{a}\right)(1-\alpha)^2\alpha\|f_{\theta^{t-2}} - f_{\theta^t}\|^2 \\
&\quad + \frac{\epsilon}{C}\left(1+\frac{1}{a}\right)(1-\alpha)^3\|f_{\theta^{t-1}} - f_{\theta^{t-2}}\|^2 \\
&\quad \text{(by } \epsilon\text{-sensitivity)} \\
&\leq \frac{\epsilon}{C}\left((1+a)\alpha^3 + 2\left(1+\frac{1}{a}\right)(1-\alpha)^2\alpha + \left(1+\frac{1}{a}\right)(1-\alpha)^3\right)m_t^2,
\end{aligned}$$

where $m_t^2 = \max\{\|f_{\theta^{t-1}} - f_{\theta^t}\|^2, \|f_{\theta^{t-2}} - f_{\theta^{t-1}}\|^2\}$.

In conclusion, we derive the following bound:

$$\|f_{\theta^{t+1}} - f_{\theta^t}\|_{p_t} \leq \frac{\sqrt{\epsilon}Mm_t}{\sqrt{C}\gamma}\sqrt{\left((1+a)\alpha^3 + 2\left(1+\frac{1}{a}\right)(1-\alpha)^2\alpha + \left(1+\frac{1}{a}\right)(1-\alpha)^3\right)}.$$

Using Assumption 2, we further obtain:

$$\begin{aligned}
C\|f_{\theta^{t+1}} - f_{\theta^t}\|_{p_t}^2 &:= C\int \|f_{\theta^{t+1}}(x) - f_{\theta^t}(x)\|^2 p_t(x)dx \\
&= \frac{C}{2}\int \|f_{\theta^{t+1}}(x) - f_{\theta^t}(x)\|^2 p_{f_{\theta^t}}(x)dx + \frac{C}{2}\int \|f_{\theta^{t+1}}(x) - f_{\theta^t}(x)\|^2 p_{f_{\theta^{t-1}}}(x)dx \\
&= \frac{C}{2}\|f_{\theta^{t+1}} - f_{\theta^t}\|_{f_{\theta^t}}^2 + \frac{C}{2}\|f_{\theta^{t+1}} - f_{\theta^t}\|_{f_{\theta^{t-1}}}^2 \\
&\geq \|f_{\theta^{t+1}} - f_{\theta^t}\|^2
\end{aligned} \tag{42}$$

Substituting this back into the previous inequality, we finally get:

$$\|f_{\theta^{t+1}} - f_{\theta^t}\| \leq \frac{\sqrt{\epsilon}Mm_t}{\gamma}\sqrt{\left((1+a)\alpha^3 + 2\left(1+\frac{1}{a}\right)(1-\alpha)^2\alpha + \left(1+\frac{1}{a}\right)(1-\alpha)^3\right)}.$$

By setting $\alpha = \frac{1}{2}$, minimizing over $a > 0$, we have:

$$\|f_{\theta^{t+1}} - f_{\theta^t}\| \leq \left(\sqrt{\frac{\sqrt{3}+2}{4}}\right)\frac{\sqrt{\epsilon}Mm_t}{\gamma}. \tag{43}$$

Convergence to a Stable Point. By expanding the max term in Equation 43 we establish the following bound:

$$\|f_{\theta^{t+1}} - f_{\theta^t}\| \leq \left(\sqrt{\frac{\sqrt{3}+2}{4}}\right)^{\lfloor \frac{t}{2} \rfloor}\left(\frac{\sqrt{\epsilon}M}{\gamma}\right)^{\lceil \frac{t}{2} \rceil}\|f_{\theta^1} - f_{\theta^0}\|.$$

Combining this inequality with Lemma 9 and assuming $\frac{\sqrt{\epsilon}M}{\gamma} \leq 4\sqrt{\frac{\sqrt{3}+2}{4}}$, we obtain:

$$\|f_{\theta^{t+1}} - f_{\theta^t}\| \leq 2\left(\sqrt{\frac{\sqrt{3}+2}{4}}\frac{\sqrt{\epsilon}M}{\gamma}\right)^{\frac{t}{2}}\|f_{\theta^1} - f_{\theta^0}\|. \tag{44}$$

For clarity, let $\alpha = \left(\sqrt{\frac{\sqrt{3}+2}{4}}\frac{\sqrt{\epsilon}M}{\gamma}\right)^{\frac{1}{2}}$, resulting in:

$$\begin{aligned}\|f_{\theta^{t+k}} - f_{\theta^t}\| &\leq \sum_{i=0}^{k-1}\|f_{\theta^{t+i+1}} - f_{\theta^{t+i}}\| \leq 2\alpha^t\|f_{\theta^1} - f_{\theta^0}\|\left(\sum_{i=0}^{k-1}\alpha^i\right) \\ &= 2\alpha^t\left(\frac{1-\alpha^{k-1}}{1-\alpha}\right)\|f_{\theta^1} - f_{\theta^0}\| \overset{\text{(assuming } \alpha<1)}{\leq} 2\left(\frac{\alpha^t}{1-\alpha}\right)\|f_{\theta^1} - f_{\theta^0}\|.\end{aligned}$$

Notice that the right-hand side of this inequality is independent of k. With $\alpha = \left(\sqrt{\frac{\sqrt{3}+2}{4}}\frac{\sqrt{\epsilon}M}{\gamma}\right)^{\frac{1}{2}} < 1$, for any $\delta > 0$, there exists $t > 1$ such that for all $m > t$, $\|f_{\theta^m} - f_{\theta^t}\| \leq \delta$. Thus, the sequence is Cauchy with respect to the norm $\|\cdot\|$; and by the compactness (and therefore completeness) of F, it converges to a point f^*.

To show that f^* is a stable point, we start by showing the continuity of the mapping

$$\mathcal{U}(f_{\theta_1}, f_{\theta_2}) = \arg\min_{f\in\mathcal{F}}\mathbb{E}_{(x,y)\sim D(f_{\theta_1}, f_{\theta_2})}\left[\ell(f(x), y)\right],$$

where $D(f_{\theta_1}, f_{\theta_2}) = \frac{D(f_{\theta_1}) + D(f_{\theta_2})}{2}$. Applying Lemma 10 and Assumption 6, we obtain:

$$\chi^2(D(f_{\theta_1}, f_{\theta_2}), D(f_{\theta_1'}, f_{\theta_2'})) \leq \chi^2(D(f_{\theta_1}), D(f_{\theta_1'})) + \chi^2(D(f_{\theta_2}), D(f_{\theta_2'})) \leq \frac{\epsilon}{C}\|f_{\theta_1} - f_{\theta_1'}\|^2 + \frac{\epsilon}{C}\|f_{\theta_2} - f_{\theta_2'}\|^2.$$

Combining this with equations 41 and 42, we derive:

$$\gamma^2\|\mathcal{U}(f_{\theta_1}, f_{\theta_2}) - \mathcal{U}(f_{\theta_1'}, f_{\theta_2'})\|^2 \leq C^2M^2\chi^2(D(f_{\theta_1}, f_{\theta_2}), D(f_{\theta_1'}, f_{\theta_2'})) \leq \epsilon M^2(\|f_{\theta_1} - f_{\theta_1'}\|^2 + \|f_{\theta_2} - f_{\theta_2'}\|^2). \tag{45}$$

Thus, for any sequence $\lim_{n\to\infty}(f_{\theta_n^1}, f_{\theta_n^2}) = (f_{\theta^1}, f_{\theta^2})$,

$$\lim_{n\to\infty} \|\mathcal{U}(f_{\theta_n^1}, f_{\theta_n^2}) - \mathcal{U}(f_{\theta^1}, f_{\theta^2})\| \leq \lim_{n\to\infty} \frac{\epsilon M^2}{\gamma^2}(\|f_{\theta_n^1} - f_{\theta^1}\|^2 + \|f_{\theta_n^2} - f_{\theta^2}\|^2) = 0.$$

This implies that if $\lim_{n\to\infty}(f_{\theta_n^1}, f_{\theta_n^2}) = (f_{\theta^1}, f_{\theta^2})$, then $\lim_{n\to\infty} \|\mathcal{U}(f_{\theta_n^1}, f_{\theta_n^2}) - \mathcal{U}(f_{\theta^1}, f_{\theta^2})\| = 0$.
By the continuity of $\mathcal{U}$, we conclude:

$$f^* = \lim_{t\to\infty} f_{\theta^{t+1}} = \lim_{t\to\infty} \mathcal{U}(f_{\theta^t}, f_{\theta^{t-1}}) = \mathcal{U}\left(\lim_{t\to\infty} f_{\theta^t}, \lim_{t\to\infty} f_{\theta^{t-1}}\right) = \mathcal{U}(f^*, f^*).$$

This establishes that $f^* = f_{\theta_{PS}}$ is a stable point.

Appendix H. Lower bound in Perdomo et al. [21] Framework

In this proof, we begin by considering a loss function defined as follows:

$$\ell(z, \theta) = \frac{\gamma}{2}\|\theta - \frac{\beta}{\gamma}z\|^2. \tag{46}$$

This function is γ-strongly convex for the parameter θ and its gradient with respect to θ is β-Lipschitz in sample space. The necessary assumptions on the loss function, as outlined in Perdomo et al. [21], are satisfied by this formulation.

We define the matrix A within $\mathbb{R}^{d\times d}$ as:

$$A = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 0 & 1 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 1 \end{bmatrix}.$$

The critical property of this matrix is that if a vector b The key property of this matrix is that if a vector $b \in \text{span}\{e_i \mid i \leq t\}$, then $Ab \in \text{span}\{e_i \mid i \leq t+1\}$, where each $e_i \in \mathbb{R}^d$ is a standard basis vector with all coordinates zero except for the i-th coordinate, which is 1. This structure enables the introduction of a new dimension only at the end of each *RRM* iteration. With the correct initialization, this ensures that the updates remain within a minimum distance from the stable point due to undiscovered dimensions.

We define $\mathcal{D}(\theta)$ as the distribution of z given by:

$$z \sim \mathcal{N}\left(\frac{\epsilon}{2}A\theta + e_1, \sigma^2\right).$$

Note that since spectral radius A is 2, the mapping $\mathcal{D}(.)$ defined as above would be ϵ-sensitive. Under this setting, the first-order Repeated Risk Minimization (RRM) update, starting with $\theta_0 = e_1$, is described by:

$$\theta^{t+1} = \frac{\beta}{\gamma}\left(\frac{\epsilon}{2}A\theta^t + e_1\right),$$

Due to the properties of matrix A, we conclude that $\theta^{t+1} \in \text{span}\{e_i \mid \forall i \leq t+1\}$.

The stationary point θ_{PS} of this setup is located at:

$$\theta_{PS} = \left(\frac{\gamma}{\beta}I - \frac{\epsilon}{2}A\right)^{-1} e_1,$$

Note that at time step t the best model within the feasible set is $\theta^t \in \text{span}\{e_i | \forall i \leq t\}$. Given that one can conclude that the best L1-distance to stationary point achievable at time step t is lower bounded by the sum over the last $d-t$ entries of θ_{PS}. Setting $d = 2T$ and using Lemma 4 we get

$$\|\theta^t - \theta_{PS}\| = \Omega\left((\frac{\epsilon\beta}{2\gamma})^t\right).$$

Similar to Repeated Risk Minimization (RRM), the Repeated Gradient Descent (RGD) method introduces a new dimension in each iteration step. Specifically, the gradient update rule in RGD is given by:

$$\mathbb{E}_{z\sim\mathcal{D}(\theta^t)}\nabla_\theta \ell(z, \theta) = \gamma\theta^t + \beta\left(\frac{\epsilon}{2}A\theta^t + e_1\right),$$

This formulation ensures that each step effectively augments the dimensionality of the parameter space being explored only by a single dimension. Consequently, the lower bound established for RRM also applies to these RGD settings.

Appendix I. Lower Bound for Mofakhami et al. [17] Framework

We define the model fitting function as $f_\theta(x) = \theta$, and the corresponding loss function is:

$$\ell(x, \theta) = \frac{1}{2\gamma} \left\| \gamma f_\theta(x) - M(1-\delta) x e^{-\frac{1}{2e}\|x\|^2} \right\|^2.$$

This loss is γ-strongly convex, ensuring unique minimizers and stable convergence properties. Additionally, we assume $\theta \in \Theta = \{z \| \|z\| \leq \frac{\delta M}{\gamma}\}$, ensuring that the gradient norm $\|\gamma\theta - M(1-\delta)xe^{-\frac{1}{2e}\|x\|^2}\|$ remains bounded by M. This holds because the mapping $f(x) = xe^{-\frac{1}{2e}\|x\|^2}$ is chosen such that, $f : \mathbb{R} \to [0, 1]$.

Observe that, for all $f_{\theta*}, f_\theta, f_{\theta'} \in \mathcal{F}$, we have $\|\theta - \theta'\| = \|f_\theta - f_{\theta'}\|_{f_{\theta*}}$:

$$\|f_\theta - f_{\theta'}\|^2_{f_{\theta*}} = \int \|f_\theta(x) - f_{\theta'}(x)\|^2 p_{f_{\theta*}}(x)\, dx = \int \|\theta - \theta'\|^2 p_{f_{\theta*}}(x)\, dx = \|\theta - \theta'\|^2.$$

We define the distribution mapping as follows:

$$D(\theta) = N\left(\sqrt{\frac{\sigma^2\epsilon}{2}} A\theta + \frac{e_1}{L}, \sigma^2 I\right),$$

where A is a lower triangular matrix:

$$A = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 0 & 1 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 1 \end{bmatrix}.$$

Matrix A has the property that if b is in the span of $\{e_1, \dots, e_i\}$, then Ab will be in the span of $\{e_1, \dots, e_{i+1}\}$. Here, e_i denotes the standard basis vector, where its i-th element is 1 and all other elements are 0. This makes A crucial for ensuring that each update step involves interactions that span progressively larger subspaces.

The χ^2-divergence between two distributions $D(\theta_1) = N(\mu_1, \Sigma)$ and $D(\theta_2) = N(\mu_2, \Sigma)$, where $\mu_1 = \sqrt{\frac{\sigma^2\epsilon}{2}} A\theta + \frac{e_1}{L}$ and $\mu_2 = \sqrt{\frac{\sigma^2\epsilon}{2}} A\theta' + \frac{e_1}{L}$, with $\Sigma = \sigma^2 I$, according to Lemma 6:

$$\chi^2(N(\mu_1, \Sigma), N(\mu_2, \Sigma)) \leq \frac{1}{2}(\mu_1 - \mu_2)^\top \Sigma^{-1}(\mu_1 - \mu_2) = \frac{1}{\sigma^2}(\sqrt{\frac{\sigma^2\epsilon}{2}} A)^2 \|\theta_1 - \theta_2\|^2.$$

Since the spectral norm of matrix A is 2, we have:

$$\chi^2(D(\theta_1), D(\theta_2)) \leq \epsilon\|\theta_1 - \theta_2\|^2 = \epsilon\|f_\theta - f_{\theta'}\|_{f_\theta}.$$

Thus, the χ^2-divergence between the distributions is bounded by $\epsilon\|\theta - \theta'\|^2$, ensuring that the divergence scales with the difference between θ and θ'.

The update rule for θ is:

$$\theta^{t+1} = \text{proj}_\Theta\left(\frac{M}{\gamma}(1-\delta)\mathbb{E}\left[xe^{-\frac{1}{2e}\|x\|^2}\right]\right) = \text{proj}_\Theta\left(\frac{M}{\gamma}(1-\delta)\exp\left(-\frac{\|\mathbb{E}[x]\|^2}{2\sigma^2}\left(1 - \frac{1}{\frac{\sigma^2}{e}+1}\right)\right) \cdot \frac{\mathbb{E}[x]}{\frac{\sigma^2}{e}+1}\right).$$

This is the unique minimizer of the loss function due to the γ-strong convexity. Additionally, this is a continuous mapping from a compact convex set Θ to itself. By Schauder 's fixed-point theorem, there exists a stable fixed point, denoted as θ_{PS}, satisfying:

$$\theta_{PS} = \text{proj}_\Theta\left(\frac{Mc_{\sigma^2,\theta_{PS}}}{\gamma}(1-\delta)\mathbb{E}[x]\right) = \text{proj}_\Theta\left(\frac{Mc_{\sigma^2,\theta_{PS}}}{\gamma}(1-\delta)\left(\sqrt{\frac{\sigma^2\epsilon}{2}}A\theta_{PS} + \frac{e_1}{L}\right)\right),$$

where $c_{\sigma^2,\theta_{PS}} = \frac{\exp\left(-\frac{\|\mathbb{E}[x]\|^2}{2\sigma^2}\left(1-\frac{1}{\frac{\sigma^2}{e}+1}\right)\right)}{\frac{\sigma^2}{e}+1} \leq 1$. Assuming $\frac{M\sqrt{\epsilon}}{\gamma} \leq 1$ and $\sigma \leq \frac{\sqrt{2}}{2}$ we get that,

$$\begin{aligned}
\|\frac{Mc_{\sigma^2,\theta_{PS}}}{\gamma}(1-\delta)\left(\sqrt{\frac{\sigma^2\epsilon}{2}}A\theta_{PS} + \frac{e_1}{L}\right)\| &\leq \|\frac{Mc_{\sigma^2,\theta_{PS}}}{\gamma}(1-\delta)\sqrt{\frac{\sigma^2\epsilon}{2}}A\theta_{PS}\| + \frac{Mc_{\sigma^2,\theta_{PS}}}{\gamma}(1-\delta)\frac{1}{L} \\
&\leq \|\frac{1}{2}(1-\delta)\theta_{PS}\| + \frac{M}{\gamma}(1-\delta)\frac{1}{L} \\
&\leq \frac{1}{2}(1-\delta)\delta + \frac{M}{\gamma}(1-\delta)\frac{1}{L}
\end{aligned}$$

Choosing $L \geq \frac{2M(1-\delta)}{\gamma(\delta+\delta^2)}$ one can guarantee the term in the projection operation would have a norm smaller than δ, i.e. it would be in Θ. So you can drop the projection operation from the equality above.

Thus, the stable point would hold true in the following equality:

$$\theta_{PS} = \left(I - (1-\delta)\frac{c_{\sigma^2,\theta_{PS}}}{\sqrt{2}}\frac{\sqrt{\sigma^2\epsilon}M}{\gamma}A\right)^{-1}\frac{e_1}{L}.$$

The same assumptions stated above would allow us to use Lemma 4:

$$\|\theta^t - \theta_{PS}\| = \Omega\left(\left((1-\delta)\frac{c_{\sigma^2,\theta_{PS}}}{\sqrt{2}}\frac{\sqrt{\sigma^2\epsilon}M}{\gamma}\right)^t\right).$$

To lower bound $c_{\sigma^2,\theta_{PS}}$, we note that $\|\mathbb{E}[x]\| \in [0, \frac{\epsilon}{2}\delta + \frac{\gamma(\delta+\delta^2)}{2M(1-\delta)}]$ and minimize the exponential term with respect to σ^2:

$$\exp\left(-\frac{\|\mathbb{E}[x]\|^2}{2\sigma^2}\left(1-\frac{1}{\frac{\sigma^2}{e}+1}\right)\right) \geq \exp\left(-c\delta\right),$$

Where $c > 0$ is a constant independent of δ. Setting $\sigma = \frac{\sqrt{2}}{2}$ to maximise $\frac{\sigma}{\frac{\sigma^2}{e}+1}$, and $\lim \delta \to 0$, we achieve:

$$\|\theta^t - \theta_{PS}\| = \Omega\left(\left(\frac{1}{\frac{1}{e}+2}\frac{\sqrt{\epsilon}M}{\gamma}\right)^t\right).$$

Hence,

$$\|f_{\theta^t} - f_{\theta_{PS}}\|_{f_{\theta_{PS}}} = \Omega\left(\left(\frac{1}{\frac{1}{e}+2}\frac{\sqrt{\epsilon}M}{\gamma}\right)^t\right).$$

Appendix J. Proof of Theorem 5

Consider a feature vector x divided into strategic features x_s and non-strategic features x_f, so that $x = (x_s, x_f)$. We resample only the strategic features with probability $g(f_\theta(x))$, representing the probability of rejection for x. The pdf of the modified distribution p_{f_θ} is:

$$p_{f_\theta}(x) = p(x)\,(1 - g(f_\theta(x))) + \int_{x'_s} p(x'_s, x_f)\, g(f_\theta(x'_s, x_f))\, p(x_s) dx'_s,$$

where the integral is over all possible values of x'_s with x_f held constant, since only the strategic features are resampled. The first term represents the option that we accept the first sample at x; the second term represents the possibility that we reject the first sample at $x' = (x'_s, x_f)$ and then resample at x_s to obtain x as well.

Assuming the strategic and non-strategic features are independent, we can rewrite this expression as:

$$\begin{aligned}
p_{f_\theta}(x) =& p(x)(1 - g(f_\theta(x))) + \int_{x'_s} p(x'_s, x'_f = x_f)\, g(f_\theta(x'))\, p(x_s) dx'_s \\
&= p(x)(1 - g(f_\theta(x))) + \int_{x'_s} g(f_\theta(x')) p_{X_s}(x'_s) p_{X_f}(x_f) p_{X_s}(x_s) dx'_s \\
&= p(x)(1 - g(f_\theta(x))) + \int_{x'_s} g(f_\theta(x')) p_{X_s}(x'_s) p(x) dx'_s \\
&= p(x)\Big((1 - g(f_\theta(x))) + \int_{x'_s} g(f_\theta(x')) p_{X_s}(x'_s) dx'_s\Big) \\
&= p(x)\,(1 - g(f_\theta(x)) + C_\theta(x_f))\,,
\end{aligned} \tag{47}$$

where p_{X_s} and p_{X_f} are the marginal distributions of the strategic and non-strategic features, respectively, and we define:

$$C_\theta(x_f) = \int_{x'_s} p_{X_s}(x'_s)\, g(f_\theta(x'_s, x_f))\, dx'_s.$$

Since $0 \le f_\theta(x) \le 1 - \delta$ for some $\delta > 0$, it follows that $\delta \le g(f_\theta(x)) \le 1$ for every x. Therefore, $\delta \le C_\theta(x_f) \le 1$.

In the RIR procedure, the distribution of the label y given x is not affected by the predictions so for every $z = (x, y)$ we have $p_{f_\theta}(z) = p_{f_\theta}(x) p(y|x)$ for any f_θ. This results in the following equality:

$$\chi^2(\mathcal{D}(f_{\theta'}), \mathcal{D}(f_\theta)) = \int \frac{(p_{f_{\theta'}}(z) - p_{f_\theta}(z))^2}{p_{f_\theta}(z)} dz = \int \frac{(p_{f_{\theta'}}(x) - p_{f_\theta}(x))^2}{p_{f_\theta}(x)} dx$$

We prove that this mapping is ϵ-sensitive with respect to χ^2 divergence, where $\epsilon = \frac{1}{\delta}\left(1 + \frac{1-\delta}{2\sqrt{\delta}}\right)$.

$$\begin{aligned}
\chi^2(\mathcal{D}(f_{\theta'}), \mathcal{D}(f_\theta)) &= \int \frac{\left(p_{f_{\theta'}}(x) - p_{f_\theta}(x)\right)^2}{p_{f_\theta}(x)}\, dx \\
&= \int \frac{p(x)^2 \left[f_\theta(x) - f_{\theta'}(x) - (C_\theta(x_f) - C_{\theta'}(x_f))\right]^2}{p(x)\left(1 - f_\theta(x) - \delta + C_\theta(x_f)\right)}\, dx \\
&\leq \frac{1}{\delta} \int p(x) \Bigg[\left(f_\theta(x) - f_{\theta'}(x)\right)^2 + \left(C_\theta(x_f) - C_{\theta'}(x_f)\right)^2 \\
&\qquad - 2\left(f_\theta(x) - f_{\theta'}(x)\right)\left(C_\theta(x_f) - C_{\theta'}(x_f)\right) \Bigg] dx
\end{aligned}$$

This inequality follows from the fact that $\delta \leq C_\theta(x_f)$ and $1 - g(f_\theta(x)) \geq 0$, therefore $\frac{1}{1-g(f_\theta(x))+C_\theta(x_f)} \leq \frac{1}{\delta}$.

Continuing, we have:

$$\begin{aligned}
&= \frac{1}{\delta} \Bigg[\int p(x) \left(f_\theta(x) - f_{\theta'}(x)\right)^2 dx \\
&\qquad + \int_{x_f} p_{X_f}(x_f) \left(C_\theta(x_f) - C_{\theta'}(x_f)\right)^2 dx_f \\
&\qquad - 2 \int_{x_f} p_{X_f}(x_f) \left(C_\theta(x_f) - C_{\theta'}(x_f)\right) \int_{x_s} p_{X_s}(x_s) \left(f_\theta(x) - f_{\theta'}(x)\right) dx_s\, dx_f \Bigg] \\
&= \frac{1}{\delta} \left[\int p(x) \left(f_\theta(x) - f_{\theta'}(x)\right)^2 dx - \int_{x_f} p_{X_f}(x_f) \left(C_\theta(x_f) - C_{\theta'}(x_f)\right)^2 dx_f \right] \\
&\leq \frac{1}{\delta} \int p(x) \left(f_\theta(x) - f_{\theta'}(x)\right)^2 dx
\end{aligned}$$

This comes from the fact that $\int_{x'_s} p_{X_s}(x'_s)(f_\theta((x'_s, x_f)) - f_{\theta'}((x'_s, x_f)))dx'_s = C_\theta(x_f) - C_{\theta'}(x_f)$. We use equation 47 to replace $p(x)$.

$$\begin{aligned}
&= \frac{1}{\delta} \int \left(p_{f_\theta}(x) + p(x)\left(f_\theta(x) + \delta - C_\theta(x_f)\right)\right)\left(f_\theta(x) - f_{\theta'}(x)\right)^2 dx \\
&= \frac{1}{\delta} \|f_\theta - f_{\theta'}\|_{f_\theta}^2 + \frac{1}{\delta} \int p(x) \left(f_\theta(x) + \delta - C_\theta(x_f)\right)\left(f_\theta(x) - f_{\theta'}(x)\right)^2 dx \\
&\overset{\text{Cauchy-Schwarz Ineq.}}{\leq} \frac{1}{\delta} \|f_\theta - f_{\theta'}\|_{f_\theta}^2 + \frac{1}{\delta} \left(\int p(x) \left(f_\theta(x) + \delta - C_\theta(x_f)\right)^2 dx \right)^{1/2} \left(\int p(x) \left(f_\theta(x) - f_{\theta'}(x)\right)^4 dx \right)^{1/2} \\
&\leq \frac{1}{\delta} \|f_\theta - f_{\theta'}\|_{f_\theta}^2 + \frac{1}{\delta} \left(\int_{x_f} p_{X_f}(x_f) Var_{x_s}\left[g(f_\theta(x))\right] dx_f \right)^{1/2} \left(\int p(x) \left(f_\theta(x) - f_{\theta'}(x)\right)^4 dx \right)^{1/2}
\end{aligned}$$

Since $g(f_\theta(x))$ is a bounded random variable in $[\delta, 1]$, its variance is less than $\frac{(1-\delta)^2}{4}$, according to Popoviciu's inequality. Also since for any $\theta \in \Theta$ we have $f_\theta(x) \leq 1$ we can infer $|f_\theta(x) - f_{\theta'}(x)| \leq$

1

$$\leq \frac{1}{\delta}\|f_\theta - f_{\theta'}\|_{f_\theta}^2 + \frac{1-\delta}{2\delta}\left(\int p(x)\left(f_\theta(x) - f_{\theta'}(x)\right)^4 dx\right)^{1/2}$$
$$\leq \frac{1}{\delta}\|f_\theta - f_{\theta'}\|_{f_\theta}^2 + \frac{1-\delta}{2\delta}\left(\int p(x)\left(f_\theta(x) - f_{\theta'}(x)\right)^2 dx\right)^{1/2}$$
$$\leq \frac{1}{\delta}\|f_\theta - f_{\theta'}\|_{f_\theta}^2 + \frac{1-\delta}{2\delta}\|f_\theta - f_{\theta'}\|$$

From Appendix A.3 in Mofakhami et al. [17], we know that $\|f_\theta - f_{\theta'}\|^2 \leq \frac{1}{\delta}\|f_\theta - f_{\theta'}\|_{f_\theta}^2$. Hence,

$$\chi^2(\mathcal{D}(f_{\theta'}), \mathcal{D}(f_\theta)) \leq \frac{1}{\delta}\left(1 + \frac{1-\delta}{2\sqrt{\delta}}\right)\|f_\theta - f_{\theta'}\|_{f_\theta}^2$$

Rate Improvement Arguments: By using Assumptions 4 and 6 from Mofakhami et al. [17], it can be shown that the method is $C\epsilon$-sensitive as defined in Assumption 1. Specifically,

$$\chi^2(\mathcal{D}(f_{\theta'}), \mathcal{D}(f_\theta)) \leq \epsilon\|f_\theta - f_{\theta'}\|^2 \leq C\epsilon\|f_\theta - f_{\theta'}\|_{f_\theta}^2.$$

In this case, our rate aligns with the rate from Mofakhami et al. [17], demonstrating that in all cases where their rate holds, our approach offers at least an equivalent or faster rate. However, there are instances where our rate results in a smaller constant than $C\epsilon$. As outlined in Appendix A.3 of Mofakhami et al. [17], the same RIR framework derives $C = \frac{1}{\delta}$ under Assumption 2 and $\epsilon = \frac{1}{\delta}$ with respect to Assumption 6, yielding $C\epsilon = \frac{1}{\delta^2}$. We show that instead of $C\epsilon = \frac{1}{\delta^2}$, we obtain $\frac{1}{\delta}\left(1 + \frac{1-\delta}{2\sqrt{\delta}}\right)$, which is strictly smaller for any $0 \leq \delta < 1$. This shows that this rate is a strict improvement over Mofakhami et al. [17].

Appendix K. Performative Risk for Credit-Scoring

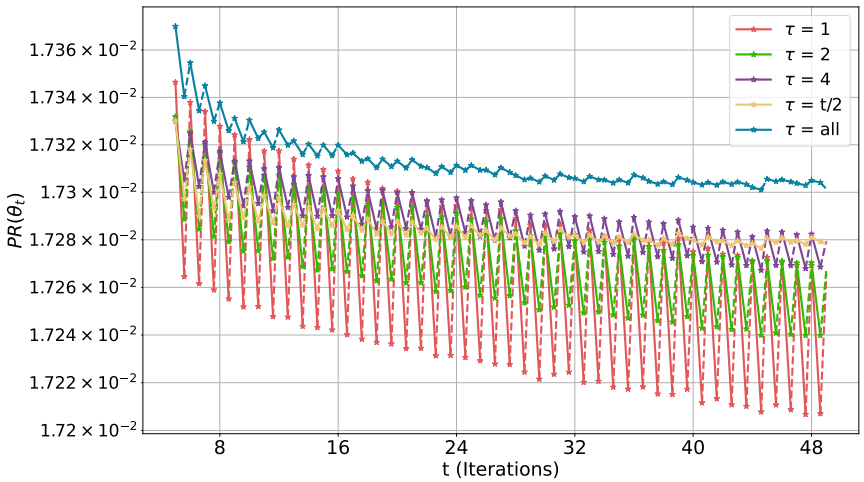

Figure 6: Log performative risk for the credit scoring environment across the RRM iterations. The numbers in the plot are averaged over 500 runs. **Increasing the size of aggregation window τ from $1 \to 2 \to 4 \to t/2 \to all$ reduces the oscillations in the risk and converges to the same point.** Note that the plot starts from iteration 5 for better readability as the initial risk values were very high.

Figure 6 shows the log performative risk for the credit-scoring environment. This metric has been adapted from Mofakhami et al. [17]. Figure 6 further substantiates our claims as we see lower oscillations in the risk for larger aggregation windows. Furthermore, another important conclusion is that most methods converge to roughly the same/very close performatively stable point as the difference in log performative risk at the end of 50 iterations is negligible between all the methods. However, as pointed out in section A, all methods oscillate in a similar range, thus hindering the readability of the plot.

Hyper-parameters. For our experiments, we fix the value of $\delta = 0.55$. The RRM procedure is carried out for a maximum of 50 iterations with a learning rate of $3e$-4 and Adam optimizer. Further, all the experimental results and plots are averaged over 500 runs, where each run for each method has the same model initialization. Thus, the only source of randomization is the sampling under *RIR* mechanism where the sampling changes across different runs but is the same for all the methods given a specific run.

Appendix L. Related Work

Performative prediction introduces a framework for learning under decision-dependent data [21], and has been widely studied in various aspects, from stochastic optimization methods to find stable classifiers [12, 14] to approaches that focus on performative optimal solutions, the minimizer of performative risk [11, 13, 16]. In this work, we focus our analysis on performative stable solutions, whose deployment removes the need for repeated retraining in changing environments [2, 10, 14].

One of the main applications of this framework is strategic classification [8] which involves deploying a classifier interacting with agents who strategically manipulate their features to alter the classifier's predictions and achieve their favorable outcomes. Strategic Classification has served as a benchmark in the literature of performative prediction [14, 16–18, 21], and we adopt this setting in our experiments to empirically demonstrate our theoretical contributions.

Prior work in performative prediction either assumes the data distribution is a function of the parameters modeled as $\mathcal{D}(\theta)$ [3, 9, 21], or more realistically dependent on the predictions as in $\mathcal{D}(f_\theta)$ [15, 17]. Although existing work only assumes one of these settings, our work adheres to both, by providing a tightness analysis of the rates proposed in Perdomo et al. [21] and Mofakhami et al. [17] and showcasing scenarios where we can provide a faster convergence rate by considering the history of distributions. To the best of our knowledge, we are first to provide a lower bound on the converge rates achievable using any such affine combination of previous snapshots.

Most related to our idea of using previous distributions are works that study gradually shifting environments considering history dependence [2, 12, 22]. Brown et al. [2] brought up the notion of stateful performative prediction studying problems where the distribution depends on the classifier and the previous state of the population. This is modeled by a transition function that is fixed but a priori unknown and they show that by imposing a Lipschitz continuity assumption similar to ϵ-sensitivity to the transition map, they can prove the convergence of RRM to an equilibrium distribution-classifier pair. In our work, we consider a specific dependence on history, by using an affine combination of previous distributions, and show that this can lead to an improved convergence than prior work without imposing any additional assumption.